Using Species Distribution Models for Spatial Conservation Planning of African Penguins

by

Frieda Geldenhuys

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematics in the Faculty of Mathematical Sciences at Stellenbosch University

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Prof. Cang Hui

Co-supervisor: Prof. Martin Nieuwoudt

March 2018

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: .....
Frieda Geldenhuys

Date: December 7, 2017 .....

Abstract

Using Species Distribution Models for Spatial Conservation Planning of African Penguins

Frieda Geldenhuys

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc. (Mathematics)

December 2017

The African penguin *Spheniscus demersus* inhabits the south-western coast of Africa, between Namibia and Algoa Bay, near Port Elizabeth, South Africa, with the largest colony consisting of about 44% of South Africa's penguins, found on St. Croix Island. The penguin population is currently at about 2% of the level it was in the 1900s, and is still continuing its strong downward population trajectory. The decrease in the population of African penguins is an early warning indicator of environmental threats, thus studying the factors that affect it is important. The African penguin has been declared Endangered on the International Union for Conservation of Nature (IUCN) Red List of Threatened Species. Due to their population decrease, immediate conservation action is required to prevent this species' extinction. An understanding of the dynamics and causes of this decrease, is thus of critical importance.

The aim of this study is to better understand the effects of drivers of change on the African penguin colonies. The establishment of a sustainable management plan for the African penguin species, by consolidating different approaches, has been investigated. Studies indicate that the drivers of change in the population size include climate change, parasites, pollution (oiling), disease, lack of food resources, predation risk and habitat interference. A large component of this is the anthropogenic impact, especially with human population expansion. As a result of this, ecological traps or scenarios in which organisms settle in habitats of poor quality, due to rapid environmental change, emerge. For example, high plankton populations could indicate high fish populations in an area, although this indicator may be incorrect if the fish have been harvested. This area may thus be an ecological trap for penguins. It is important, for conservation purposes, to be able to identify the ecological traps and differentiate them from sinks or low quality habitats that, on their own, would not have the resources

to support a population. The information required to assess the consequences of ecological traps was investigated.

Of particular concern are the shifting distributions of forage fish, which may result in a spatial mismatch between the main penguin breeding colonies and their preferred prey. The foraging range of penguins during the breeding season is particularly limited, as foraging trips typically last less than one day. Spatial closures, in the form of marine protected areas, as well as those that permanently prohibit fishing, termed no-take reserves, can be used to manage the fishing effort, and in complementing alternative controls such as quota management.

Species Distribution Models (SDMs) have been established in response to these challenges. These are predictive, conceptual models of the abiotic (e.g. physical barriers, climate, lack of resources) and biotic (e.g. competition, predators, parasites) factors influencing the role of habitat suitability in affecting the distribution of species in terms of space, time and scale.

To begin with, the demography of the African penguin has been investigated. Thereafter, the modelling method has been described. R statistical programming language has been used to create the SDMs, from the colony location inputs and corresponding environmental data. The Maximum Entropy algorithm used 5 environmental, non-correlated variables and presence-only records (from 33 colonies). The relative contributions of environmental variables, which are ecologically relevant to the species habitat suitability, indicate that sea surface temperature is the largest contributing factor, with 72.4% for annual, 53.2% for summer and 46.9% for winter factors. The second largest contributor for all seasons is mean land temperature.

The outputs of this study act as a baseline assessment. Possible areas to relocate or establish African penguin colonies, based on their prey availability, include the old De Hoop colony (which went extinct in 2006) and a site near Plettenberg Bay (which would be a completely new site), according to BirdLife. Camera traps for checking predators, have been in place since November 2016. From this study, it is clear that ongoing research is necessary, mainly due to the shifting distribution of prey, which is caused by climate change and overfishing, in order to model the African penguin colonies.

Keywords: African penguins, Conservation, Species Distribution Models, Suitability Maps

Opsomming

Gebruikmaking van Spesies Verspreiding Modelle vir Ruimtelike Bewaringsbeplanning vir Afrika Pikkewyne

Frieda Geldenhuys
Departement Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.
Tesis: MSc. (Wiskunde)
Desember 2017

Die Afrika pikkewyn *Spheniscus demersus* bewoon die suid-westelike kus van Afrika, tussen Namibië en Algoa Baai, naby Port Elizabeth, Suid-Afrika, waar die grootste kolonie bestaan uit omtrent 44% van Suid Afrika se pikkewyne, te vinde by St. Croix Eiland. Die pikkewyn bevolking is tans ongeveer 2% van die vlak wat dit was in die 1900s, en is steeds op 'n sterk afwaartse bevolkingstrajek. Die afname van die bevolking Afrika pikkewyne is 'n vroeë waarskuwings-aanwyser van omgewings-bedreigings. Dus is die bestudering van faktore wat dit beïnvloed baie belangrik. Die Afrika pikkewyn is nou geklassifiseer as Bedreig op die Internasionale Unie vir die Bewaring van die Natuur (IUBN) Rooi Lys van Bedreigde Spesies. Weens hul bevolkingsafname word onmiddellike bewarings-aksies vereis om hierdie spesie se uitsterwing te verhoed. 'n Begrip van die dinamika en oorsake van hierdie afname, is dus van kritieke belang.

Die doel van die studie is om die uitwerking van die aandrywers van verandering op die Afrika pikkewyn kolonies beter te verstaan. Die vestiging van 'n volhoubare bestuursplan vir die Afrika pikkewyn spesie, deur van verskillende benaderings gebruik te maak, is ondersoek. Studies dui daarop dat die aandrywers van hierdie verandering in bevolkingsgrootte, insluit klimaatsverandering, parasiete, besoedeling (met olie), siekte, gebrek aan voedselbronne, roofdier vyande risiko en habitat inmenging. 'n Groot komponent hiervan is die antropogeniese impak, veral met die menslike bevolkingsaanwas. As gevolg hiervan, ontstaan ekologiese slagysters of scenarios waar organismes gaan bly in habitats wat van swak gehalte is, weens die vinnige omgewingsverandering. Byvoorbeeld, hoë plankton bevolkings kan 'n aanwyser wees dat daar hoë visbevolkings in 'n spesifieke area behoort te wees, maar hierdie aanwyser kan verkeerd wees as die vis grootliks ge-oes is. So 'n gebied kan dus 'n ekologiese slagyster vir pikkewyne wees. Dit is belangrik vir bewaringsdoeleindes, om in staat te wees om ekologiese slagysters te identifiseer en om hul te onderskei van sinkgate of

lae gehalte habitats wat, op hul eie, nie die hulpbronne sou hê om 'n bevolking te onderhou nie. Die informasie was benodig word om die gevolge van die ekologiese slagysters te evalueer, is bestudeer.

Van besondere belang is die veranderende verspreiding van prooivis, wat tot gevolg kan hê dat daar 'n verkeerde ruimtelike paring is tussen die hoof pikkewyn broeikolonies en hul voorkeur prooi. Die jag reikwydte van pikkewyne gedurende die broeiseioen is besonder beperk, aangesien jag uitstappies tipies korter as een dag is. Ruimtelike sluitings, in die vorm van mariene beskermde areas, sowel as daardie gebiede wat visvangs permanent verbied, genoem geen-vangs reservate, kan gebruik word om visvangpogings te bestuur, wat dan alternatiewe beheermatreëls soos voorgeskrewe kwotas kan aanvul.

Spesies Distribusie Modelle (SDMs) is opgestel in reaksie op hierdie uitdagings. Hierdie is voorspellende, konseptuele modelle van die abiotiese (bv. fisieke versperrings, klimaat, gebrek aan bronne) en biotiese (bv. kompetisie, roofvyande, parasiete) faktore wat die rol van habitat geskiktheid beïnvloed deur die verspreiding van die spesies te raak in terme van ruimte, tyd en skaal.

Om mee te begin, word die demografie van die Afrika pikkewyn ondersoek. Daarna word die modelleringsmetode beskryf. R statistiese programmeringstaal gebruik om die SDMs te skep, vanuit die kolonie ligging invoere en ooreenkomstige omgewingsdata. Die Maksimum Entropie algoritme gebruik 5 omgewing, nie-korrelerende veranderlikes en teenwoordigheid-alleen rekords (van 33 kolonies). Die relatiewe bydraes van omgewingsveranderlikes, wat ekologies relevant is tot die spesie habitat geskiktheid, dui aan dat see oppervlak temperatuur die grootste bydraende faktor, met 72.4% vir jaarliks, 53.2% vir somer en 46.9% vir winter faktore is. Die tweede grootste bydraer vir alle seisoene is gemiddelde landstemperature.

Die resultate van die studie kan beskou word as 'n basislyn studie. Moontlike areas wat ondersoek word om die Afrika pikkewyn kolonies te verskuif of vestig, gebaseer op hul prooi beskikbaarheid, is die ou De Hoop kolonie (wat in 2006 uitgesterf het) en 'n area naby Plettenbergbaai (wat 'n totaal nuwe area sal wees), volgens BirdLife. Kamera lokvalle om die predatore te kontroleer is al geplaas vanaf November 2016. Uit hierdie studie sien ek dat deurlopende navorsing benodig word, veral as gevolg van die veranderende verspreiding van hul prooivis, wat veroorsaak word deur klimaatsverandering en oorbevissing, om die Afrika pikkewyn kolonies te modelleer.

Sleutelwoorde: Afrika pikkewyne, Bewaring, Spesies Verspreiding Modelle, Geskiktheidslandkaarte

Acknowledgements

Thank you to the Creator for biodiversity. Growing up in beautiful Cape Town, the ocean and mountains give me constant inspiration and interest, thus considering the topic of conserving the biodiversity and looking after our environmental health was not a difficult choice.

I would like to express my sincerest gratitude to my supervisor, Prof. Cang Hui, who gave me this interesting topic choice. I would also like to thank my co-supervisor Prof. Martin Nieuwoudt who inspired me to take up scuba diving. I am thankful for all their guidance, inspiration and support throughout my studies. I would like to thank Dr. Vernon Visser from UCT, Department Statistics, Ecology and Environment (SEEC) for his knowledge transfer on Species Distribution Modelling in R statistical programming. I would also like to thank my bursary holder, SACEMA, for the funding, constant interest and support. Also my family, friends and Bible-study group for their guidance. If I had to mention names the list would be too long. However, I would like to acknowledge my dad and boyfriend for their constant support throughout my studies, including listening to my presentations over and over again.

I am fortunate to have attended three conferences for my MSc. I presented my MSc work at the BioMath 2017 International Conference on Mathematical Methods and Models in Biosciences at Skukuza camp, Kruger Park during June 2017. Also, at the 58^{th} South African Statistical Association Conference in November 2016. I presented a poster of my work at the 9^{th} International Penguin congress in September 2016. I would like to thank everyone (The Southern African Foundation for the Conservation of Coastal Birds (SANCCOB), Department of Environmental Affairs (DEA), colleagues from universities and others) for their valuable comments at the conference. Here, I heard about Waddle 9-13 May 2017 and enjoyed walking 130 kilometres, from the African Penguin and Seabird Sanctuary in Gansbaai to Boulders beach in the Western Cape, to raise awareness of the decline of the African penguin and environmental matters, encouraging the public to make penguin promises (www.penguinpromises.com). We visited Stony Point and Boulders beach colonies, which were also highlights. I was also fortunate to go to the workshop on Research Data Science at the International Centre for Theoretical Physics in Trieste, Italy in August 2016. Here, I improved my computational skills by learning amongst other things about The Linux Shell, Git and GitHub, R Statistical Programming, SQL, Machine Learning and Recommender Systems, Data Visualisation and Open Science. I am very grateful for this opportunity that helped my research and will definitely help my future research too. I also enjoy teaching others at software and data carpentry workshops using these skills.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"When we save birds from large-scale threats, we see that what's good for the birds is also good for us. This is true about agriculture, fishing, and climate change. As we solve their problems, we solve ours. This is about everyone's quality of life." - Gary Langham, National Audubon Science Director.

Species, such as the African penguin, also known as *Spheniscus demersus*, are important as they play the role of an early warning system for environmental threats. By global standards, a population is considered unhealthy, and in danger, if it decreases to 10 percent of the former or pre-exploitation levels. The African penguin population is currently at about 2% of its 1900s level, 14% of its 1950s level when the first official census was conducted, and is still on a strong downward population trajectory.

African penguins are endemic to Southern Africa, breeding only in South Africa and Namibia. It is Africa's only extant penguin, other than the four species which breed at South Africa's Prince Edward Islands in the south-west Indian Ocean (Department Environmental Affairs, 2015). Figure 1.1 clearly depicts the three distinct population areas: Namibia, Western Cape and Algoa Bay.

There are about 17 000 breeding pairs left in South Africa according to Department Environmental Affairs, 2013 data. According to this data, St. Croix Island hosts the most penguins (44.35%), then Dassen Island at 15.25%, Stony Point at 11.78%, then Robben Island (7.90%) and Dyer Island (7.24%) colony. The most recent data for Namibia indicate that in 2015, there were about 5 700 to 5 800 pairs according to the Ministry of Fisheries and Marine Resources, unpublished data. A few islands have not been counted for several years (J. Kemper), creating the uncertainty of the numbers. It can thus be said, as stated by the Southern African Foundation for the Conservation of Coastal Birds (SANCCOB), that "there are less than 23 000 breeding pairs in the wild", taking into consideration the numbers of South Africa and Namibia.

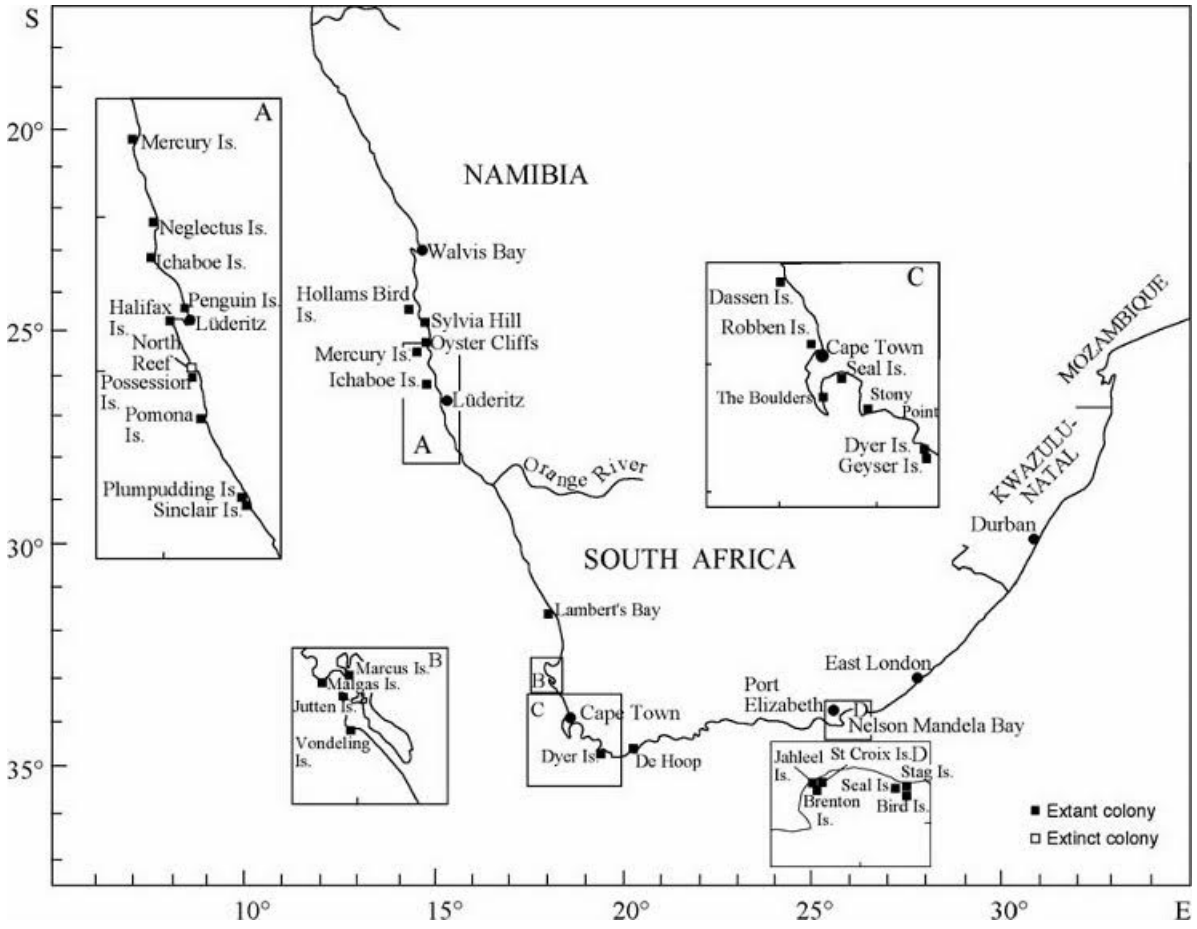

Figure 1.1: Location of African penguin colonies on the coast of South Africa. Black dots indicate current colonies. Inset shows the location of the Dassen Island and Robben Island colonies (Source: Weller et al.).

African penguins have been heading towards extinction since industrial fishing started around the Cape. The species avoids highly modified areas. In view of the fact that the downward trend in African penguin numbers currently shows no sign of reversing, immediate conservation action is required to prevent a further decline. The establishment of a sustainable management plan for the African penguin colonies, by consolidating different approaches, will be investigated. This study will use parameters to simulate the spatial and temporal variability drivers, relevant to the conservation of the African penguin. These parameters are the environmental variables used to assess the habitat quality.

Section 1.6 describes the research objectives and questions for this thesis. The Department of Environmental Affairs Biodiversity Management Plan (2013) objective 4.1.4 is: "To secure the protected status of all extant African Penguin colonies, including those not currently formally protected, and to consider the establishment of new breeding sites." This thesis will explain strategies to assist in this objective, by using Species Distribution Models (SDMs).

1.1 Life History of the African Penguin

Some of the earliest known penguin fossils have been discovered in Peru, including the 80cm tall *Perudyptes devriesi* which inhabited the Earth 42 million years ago. The more recent 150cm tall *Icadyptes salasi*, was dated as 36 million years old, when discovered (https://www.livescience.com/4518-giant-ancient-penguins-hot.html). The first animal referred to as "penguin" was a flightless bird of the Arctic ocean. It was very similar to what is now considered a penguin in terms of anatomy, however it was from a different order of birds. It was hunted to extinction in the 1600s. Later, when explorers discovered similar birds in the south seas, they gave them the same name. The word, "penguin", originally seemed to mean "fat one" in Spanish / Portuguese. It may come from either the Welsh "pen gwyn" (white head), from the Latin "pinguis" (fat) or from the derivation of "pin-wing" (pinioned wings) (https://www.penguinscience.com/education/ask-answers-6.php).

There are currently 17 species of penguins, although some scientists divide them into 18, or even 19, species. Fossil records indicate that there used to be more in the past. Currently, fifty-five percent of penguin species are considered threatened with extinction, placed as Endangered or Vulnerable on the International Union for Conservation of Nature (IUCN) status criteria (Evaluating the status and trends of Penguin Populations, Boersma et. al.). The current ones, all living in the Southern hemisphere, are: Adelie, African, Chinstrap, Emperor, Erect Crested, Fairy, Fjordland, Galapagos, Gentoo, Humboldt, King, Little, Macaroni, Magellanic, Rockhopper, Royal, Snares Island and Yellow Eyed. Some have multiple names. The current species are divided into 6 genera: *Aptenodytes, Eudyptes, Eudyptula, Megadyptes, Pygoscelis* and *Spheniscus*. More species are being discovered, however not living species. In 2008 New-Zealand researchers announced the discovery of bones belonging to a previously unknown species, the Waitaha penguin. This species went extinct about 500 years ago, soon after the human settlement of the islands.

Figure 1.2: Group of African penguins, taken by Frieda Geldenhuys whilst assisting a Phd student on site.

The African penguin's taxonomic description shows that no subspecies is recognised. The species is one of four in the genus *Spheniscus*. The current classification of *S. demersus* is as follows (Hockey et al. 2005): Order: Ciconiiformes; Family: Spheniscidae; Genus: *Spheniscus*; Species: *demersus* (Linnaeus 1758).

The genus name *Spheniscus* is derived from the ancient Greek word "sphen" (South African National Biodiversity Institute (SANBI)). This means "wedge", referring to the streamlined body shape of the African penguin. The species name *demersus* is Latin meaning plunging or sinking, and refers to its diving behaviour. The common name "jackass" refers to its braying call. It sounds similar to that of a donkey, however, most other penguins produce a similar sound, thereby giving them more distinctive names, such as the African penguin was used after 1995. Some common names for the species are: Jackass penguin, African penguin, Cape penguin, Black-Footed Penguin, Pikkewyn (Afrikaans) and Nombombiyane (Xhosa).

African penguins are flightless aquatic birds which are streamlined with reduced wings that are modified to form efficient flippers for swimming. They have heavy bones to enable them to dive. Their thick coat with overlapping feathers assists with waterproofing, wind resistance and insulation. The dorsal or back part of the body is black, and the belly is white. The white belly has a thick black stripe curving across the top of the chest, also down the flanks, towards the legs. The bare black facial mask, with distinctive pink patches of skin above the eyes aids the birds with heat regulation (Williams, 1995). To distinguish individuals from each other, each African penguin has a unique and distinct pattern of black spots on the white chest. The African penguin has a black bill, black webbed feet and shortened tail.

The colours of the penguins make them less visible when in the water. From above, only their black backs are visible above the darkness of the deep sea, whereas from below you see a light belly in front of the bright sky. They are not easily visible, either way. Many fish also have this colouration pattern. In other words, it is a defence mechanism when underwater.

The average lifespan of an African penguin is 10 to 27 years in the wild, however they can live up to the age of 30 in captivity. This beign said, there are exceptions, such as on 4 July 2017 the uShaka Sea World's beloved penguin Deé, believed to be the world's oldest African penguin, died at the age of 40 years.

African penguins are incredibly sociable birds. Adults mainly form pair bonds that last for life (as long as 10 years, see Chapter 3.2: Breeding). African penguins can often be seen grooming one another, which is not only practical for cleaning purposes and rearranging feathers, but also for removing parasites. They are constantly strengthening the social bond between the pair. It is difficult to differentiate between sexes, as males and females have the same plumage. Males can be distinguished from females by a slightly broader and bigger bill. Adults weigh on average 2.2 to 3.5 kg. They are 60 to 70 cm in height. Juveniles differ from adults in having blue-grey plumage. They have no white facial markings and no bold, delineated markings. They have dark upper-parts lacking both band and spots on the chest. Figure 1.3 shows a picture of an African penguin chick.

Figure 1.3: African penguin chick, picture taken during Waddle 2017 by Devon Bowen from Two Oceans Aquarium

African penguins are very clumsy on land. They waddle upright with flippers held away from their body as if they are drunk. They are highly specialised for a life at sea and they are efficient swimmers. Penguins can reach speeds of up to 20 km/h, cruise at 4-7 km/h and dive down to 130 m.

1.2 Conservation Status of the African Penguin

About 100 years ago, the African penguin colony at Dassen Island alone stood at about 1 million pairs (Birdlife South Africa). They were already subject to huge egg harvesting pressures and other disturbances. In 2011, around 4 000 pairs bred there. That amounts to a loss of over 10 000 pairs per year. In South Africa there are about 17 000 breeding pairs left (Department Environmental Affairs (DEA) 2013). The current global population remainder is now, at the end of the 20^{th} century, about 2% of what it was in the 1900s. African penguin populations have declined by about 98 percent since pre-industrial times. As can be seen from Figure 1.4, the last four years have seen a strong downward trajectory in the population of African penguins. The population has decreased by more than 50% in the past 30 years, signalling a strong warning to conservationists.

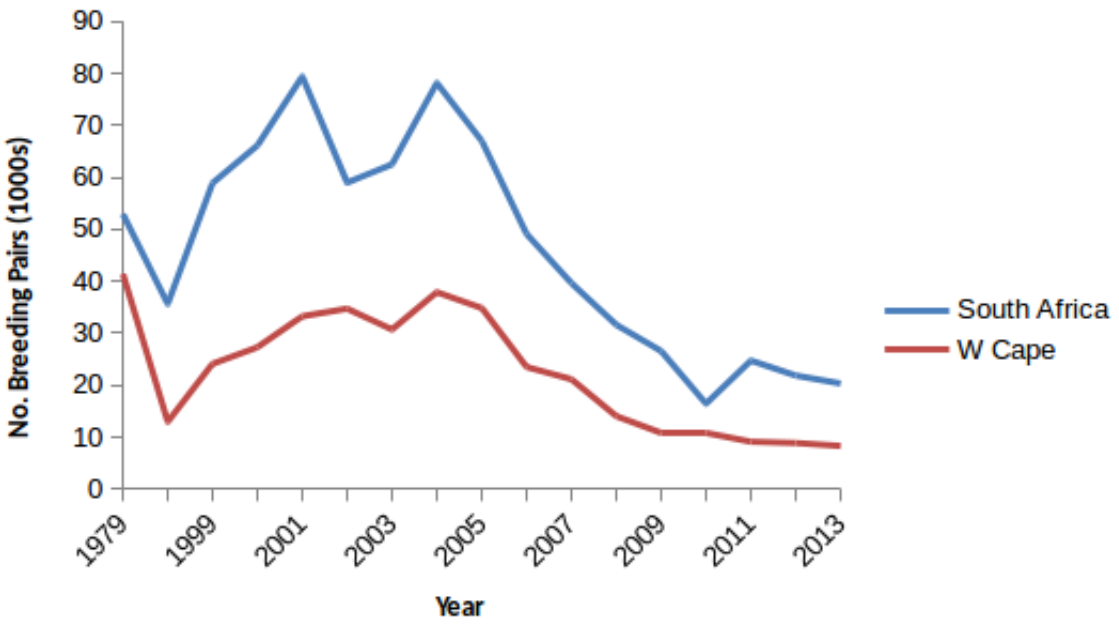

Figure 1.4: Indication of downward trajectory of the African penguin (Source: Department of Environmental Affairs (DEA)).

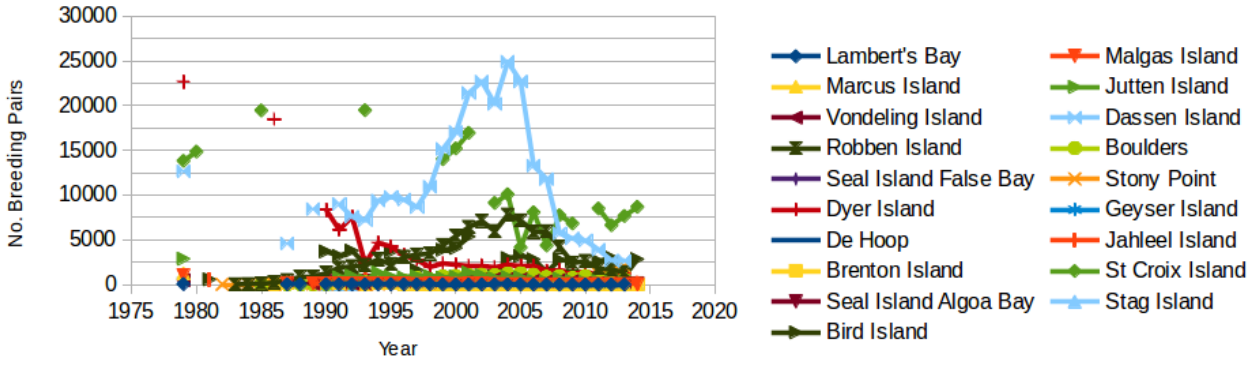

Figure 1.5: Numbers of African penguins at different colonies (Source: extracted from DEA data).

The species faces numerous threats, but the current likely drivers of the decline are food scarcity resulting from shifts in prey populations. This is possibly driven by environmental change, and competition with fisheries for prey. Due to these factors, from Figure 1.5 one can see there used to be many penguins on Dassen Island (about 25 000 pairs, 2005), but from what is left, most penguins nowadays occur on St. Croix Island.

BirdLife International has changed African penguins' conservation status from Vulnerable to Endangered, on the International Union for Conservation of Nature (IUCN) Red List of Threatened Species. The reason for this is because they have undergone this population decline of > 50%, as discussed, in the three most recent generations (Kemper 2015, Hagen 2016). The IUCN assessment is based on

rigorous criteria. A taxon is Endangered when the best available evidence indicates that it meets any of the criteria A to E for Endangered, described in Section 1.2.1, and it is therefore considered to be facing a very high risk of extinction in the wild. In a situation in which there is limited information, the data that are available can be used to provide an estimate of extinction risk. For instance, estimating the impact of stochastic events on habitat.

1.2.1 IUCN Criteria for Endangered Species

Criteria A to E briefly described $(https://en.wikipedia.org/wiki/Endangered-species)$:

A) Reduction in population size. This is based on, for example, an observed, estimated, inferred or suspected population size reduction of bigger or equal to 70% over the last 10 years, where the reduction is established, for example, by direct observation of a decline in the area of occupancy, extent of occurrence or quality of habitat.

B) Geographic range reduction, in the form of either extent of occurrence or area of occupancy, or both.

C) Population estimated to a number fewer than 2 500 mature individuals and other decline criteria. This decline criteria, could include, for example, an estimated continuing decline of at least 20% within five years or two generations, whichever is longer.

D) Population size number estimated to be fewer than 250 mature individuals.

E) Quantitative analysis showing the probability of extinction in the wild is at least 20% within 20 years or five generations, whichever is the longer (up to a maximum of 100 years).

The criteria is defined as any form of analysis which estimates the extinction probability of a taxon based on known life history, habitat requirements, threats and any specified management options discussed later. Population Viability Analysis (PVA) is one such technique.

Species that are near-critically endangered, particularly sensitive to poaching levels, near-endangered due to poaching, may vary according to levels of tourism. In particular, variation in female populations should be investigated.

In presenting the results of the quantitative analyses, the assumptions (which must be appropriate and defensible), the data used, and the uncertainty in the data or quantitative model, must be documented.

1.3 Stressors on African Penguin Populations

Contributing factors towards African penguin numbers in marine and terrestrial biodiversity, taking into consideration environmental variability, will be investigated.

The relative impact of the following factors, including human induced activities, will be studied.

Historic factors include that penguins were exploited for human consumption: the meat was pickled for sailors and large scale egg harvesting, they were rendered down for fat and used for ship fuel, and their guano (the preferred substrate for constructing nesting burrows by the penguins) scrapings were collected to be used as fertilizer.

Current factors are mainly human disturbances: tourism, poaching, habitat modification, pollution (i.e. oil spillages), overfishing (competition with commercial fishing for food resources), climatic conditions (e.g. heat stress on land and sea), causing breeding failure, introduced and natural terrestrial and marine predators, such as seals and sharks preying on adults, gulls taking eggs, as well as the effect of parasites on the health status and nesting behaviour - a PhD study is in progress (Marcela Paz A. Espinaze Pardo). These are shown in Figure 1.6.

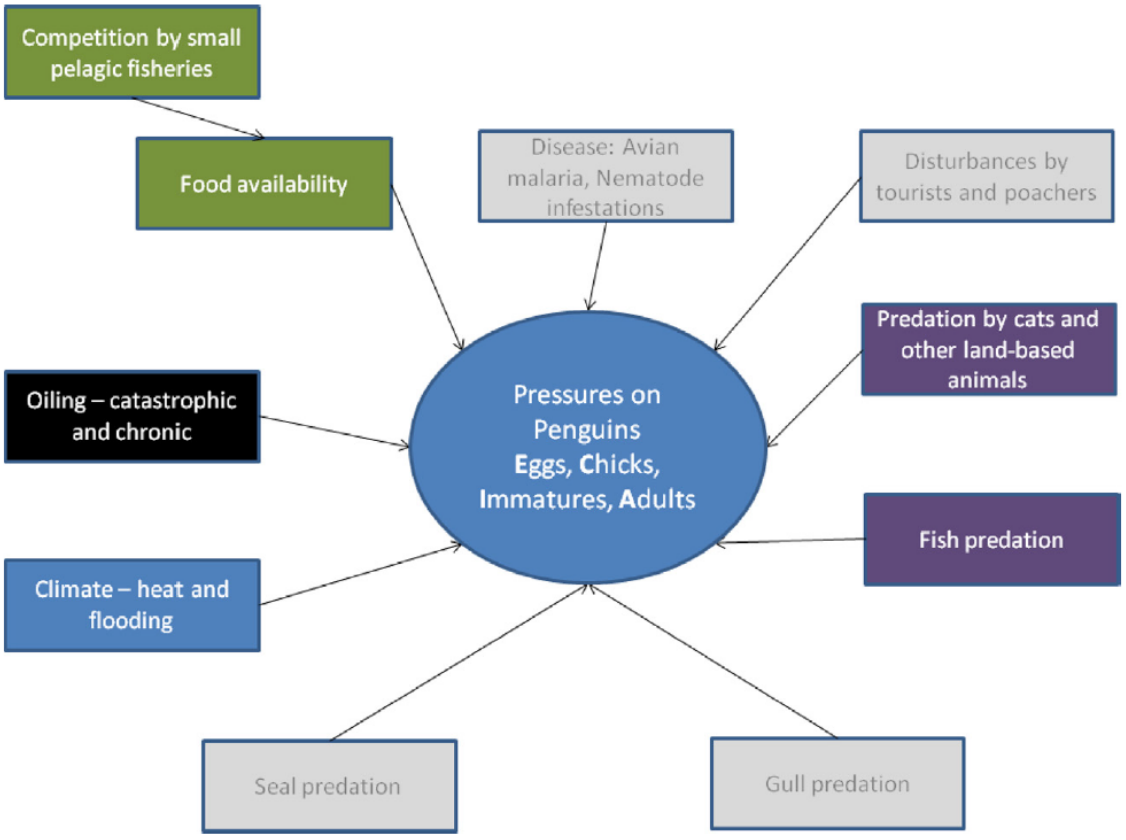

Figure 1.6: Pressures acting on penguin populations (Source: F Weller et al.).

It is not only major spills that have an impact on this species. Chronic oiling through oil from leaking containers, or through the illegal practice of ships cleaning their bilges out at sea, result in a number of penguins being oiled each year (Parsons and Underhill 2005).

Makhado (2009) documented the extent of Cape fur seal predation on South African breeding seabirds. This is considered a source of seabird mortality which is unsustainable at some colonies. The great white shark *Carcharodon carcharias* is known to predate on African penguins (Johnson et al. 2006). The

number of Kelp Gulls at some colonies has increased steadily. It is a source of predation pressure of African penguin eggs and small chicks (Kemper et al. 2007a).

A particular challenge in ecosystem modelling, which is inherently characterized by complexity and associated uncertainty, is to take the effects of climate change into account (Rose et al. 2010; Plag ányi et al. 2011a,b). Such modelling requires additional flexibility to allow for changing baselines, and adaptive management responses provide robustness to non-linear effects.

Fishing has drastically decreased sardine and anchovy populations in Namibia and western South Africa, which has cold surface waters and high chlorophyll levels, which are normally indicative of a healthy fish population. Climate change has caused the remaining fish to move southward (Crawford et al. 2017). The African penguin eats almost nothing but small pelagic fish, so when their numbers are in a steep decline, it means that there are not enough small pelagic fish for the ecosystem. Such changes not only negatively impact upon the penguins, but on the entire ecosystem, because everything else in the ecosystem relies, either directly or indirectly, on the small pelagic fish.

1.4 Challenges Facing African Penguin Conservation

In recent years, the main challenges affecting the size of the colonies include commercial fishing, marine pollution, habitat destruction and climate change. Especially of growing concern is the intensive fishing that is degrading marine ecosystems to a degree which is not sustainable. This may be driven by environmental change (Crawford et al. 2015) and competition with fisheries for prey (Crawford et al. 2011). Penguins need to cope with the heterogeneous ocean landscape, low prey availability and often long commuting between foraging and breeding areas. Human population expansion and the anthropogenic impact plays a huge role.

The large-scale collection of guano deposits along the coasts of Southern Africa, that was used as fertilizer since the mid-nineteenth century, has removed much of the breeding habitat of the penguins. This resulted in the birds breeding in a variety of suboptimal habitats (Frost et al. 1976b; Wilson and Wilson 1989). Nests are built by all penguins in burrows in guano or sand. Also, in clefts between rocks, in disused buildings and on the surface, preferably under shade (Shelton et al. 1984, Crawford et al. 1995a). Burrows have a more constant microclimate than surface nests. Relative humidity is higher, air temperatures fluctuate less, wind effect is negligible and birds are not exposed to direct sunlight (Frost et al. 1976a). Nesting material includes pieces of vegetation, seaweed, rocks, shells, bones and feathers, but some nests have no lining. As the penguins are equipped to forage in cold water, they can become heat stressed on land (Frost et al. 1976a). They breed more successfully in nest sites with cover, relative to those in the open (e.g. Frost et al. 1976b; Seddon and van Heezik 1991).

Penguins are also limited by the availability of island habitats and mainland habitats that are free from predators. There is a lack of suitable alternative sites on the Southern African coast line. Anthropogenic actions may have contributed to the decline of colonies in the past, e.g. the construction

of a land-bridge and renovation of buildings at Bird Island, Lambert's Bay and a breakwater at Marcus Island (Department Environmental Affairs (DEA)). At the colony scale, nesting habitat has been removed or degraded at a number of colonies, causing birds to nest on the surface in some cases, or to utilise lower quality nesting habitat (e.g. vegetation). Surface nesting birds are susceptible to heat stress and flooding, as well as more likely to suffer predation (both marine and terrestrial). Surface nesting may have also rendered birds more susceptible to displacement (e.g. by seals) and disturbance (e.g. by humans). Guano scraping is still a threat at some colonies in Namibia. Other disturbance to birds on land, which may cause increased stress, abandonment of chicks and/or eggs, destruction of nests and impacts on survival, usually results from direct human presence in the colony. This is due to, amongst others, research, filming, eco-tourism and poaching. Fire and vehicle strikes are potential threats at specific colonies and also need to be considered. At sea impacts on penguins include those that interfere with foraging behaviour or directly influence behaviour at sea - for example boat strikes, ghost nets and incidental by-catch of birds in fishing operations.

Some of the other challenges include the prevention and control of invasive species. Also of importance is the elimination of illegal fisheries and measures to make legal fisheries more sustainable. The control of both land-based and ship-based tourism should be looked into. The control for immigration and residency and uncontrolled human population explosion is also of importance. The measures to develop local capacity through improved education, greater transparency, accountability and efficiency in governance and regional planning are also concerns. There should be looked into the control of pollution and the protection of habitat, maintenance of biodiversity, genetic variability, and trophic level balance (Gislason et al. 2000), as well as various biological and socio-economic considerations involved in the implementation of ecosystem-based management.

Looking at the socio-economic factors in this research, we find, for example, that most colonies of African penguins are inaccessible to the general public. Two mainland colonies (Boulders and Stony Point), however, provide opportunities for the public to observe African penguins in their natural habitat. They have become popular tourist destinations. The economic benefits of these colonies include the provision of income through gate fees, provision of jobs at the colonies, as well as associated tourism benefits to the surrounding areas. Negative interactions with neighbours to these areas, as well as the risk of penguins being killed by road traffic, is managed by the relevant authorities.

At Stony Point, the number of visitors to the colony increased from 42 870 in 2008 to 69 068 in 2010. Over 10 000 visitors to the colony were recorded in December 2010 (McGeorge). The Boulders colony in Simon's Town has about 500 000 visitors annually (M Ruthenberg).

Anthropogenic climate change is recognized as a major threat to global biodiversity. The ability to predict species' responses to rapid shifts in abiotic conditions, has emerged as a conservation priority (Bellard et al., 2012; Cahill et al., 2013). There are basically two overriding factors for the choice of methods for estimating climate change vulnerability. These are: the global scale at which climate change is occurring, that means very large numbers of species must be evaluated; and the need to develop conservation interventions quickly, given accelerating rates of environmental change. Mod-

elling the distribution of species in future climates is by far the most useful means of determining how climate change will influence life on Earth (Kearney et al., 2010). In large part, this is because models can be applied rapidly to diverse taxa over large spatial scales (Pacifici et al., 2015). Use of species distribution modelling within the context of climate change and conservation research have increased in recent years.

Of particular concern are the shifting distributions of forage fish, which may result in a spatial mismatch between the main penguin breeding colonies and their preferred prey (Crawford et al. 1990; Crawford 1998). The foraging range of penguins during the breeding season is particularly limited, as foraging trips typically last less than one day (Petersen et al. 2006; Pichegru et al. 2009).

Temporal and spatial management have often been proposed as management tools that can provide an insurance against inaccuracies in stock assessments, or unknown impacts of a fishery on other species in the ecosystem. Spatial closures, such as marine protected areas, or those that permanently prohibit fishing, termed no-take reserves, can be used to manage fishing effort, complementing alternative controls such as quota management (Mangel 2000). Current Marine Protected Areas (MPAs) are shown in Figure 1.7.

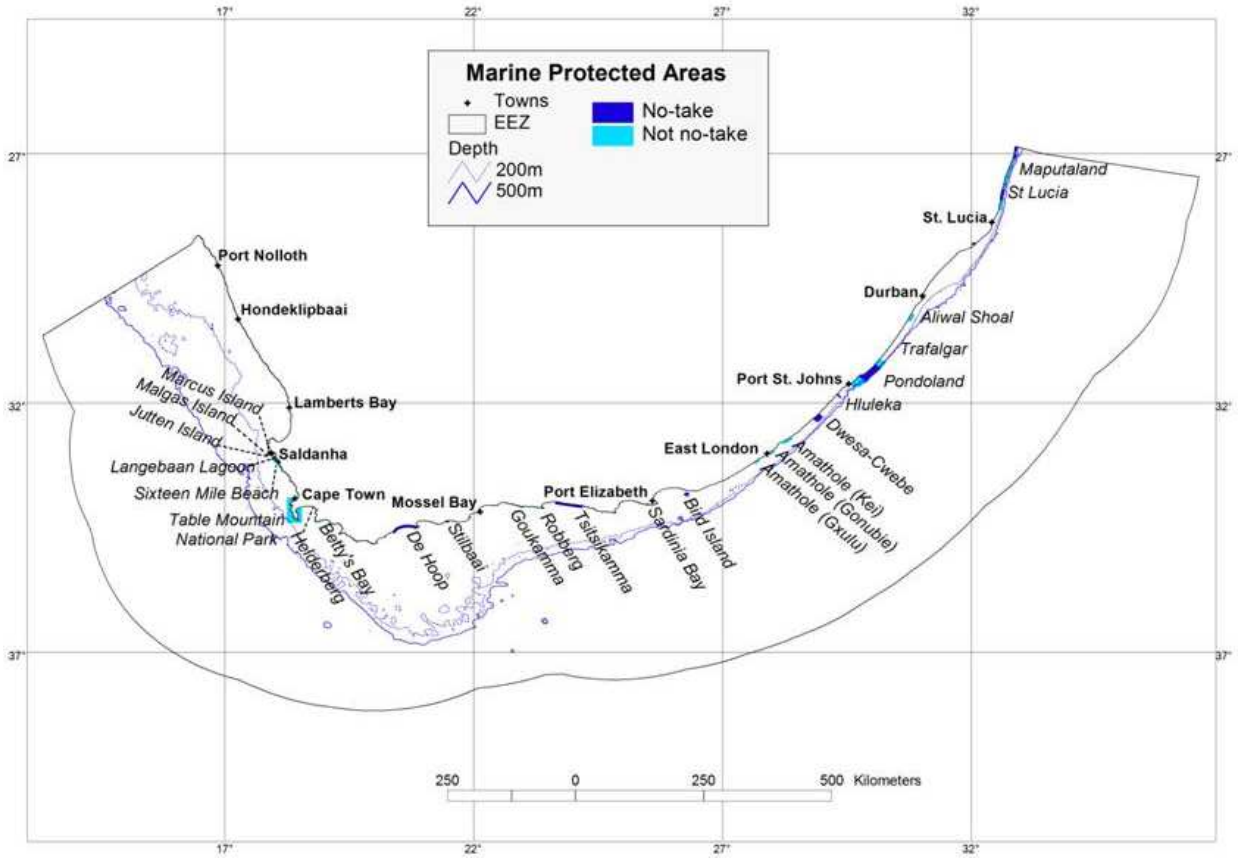

Figure 1.7: Marine Protected Areas (Source: South African National Biodiversity Institute)

Several important new features for conservation, such as a spatial aspect to sardine assessment and management are now being considered (de Moor and Butterworth 2013a), evaluating the consequences of different fishing efforts on the west and south coasts.

An important aspect, creating quite a few challenges to marine protection development is the involvement of all stakeholders, including Department Environmental Affairs (DEA) and government. Hilborn (1992) says, "Fisheries management is primarily a problem in managing people, not fish." Butterworth (2007) concludes by stating: "Industry, conservationists, scientists, and managers need to agree on the rules before a fisheries management game is played." This promotes transparency and confidence in the decision-making process, thus allowing all parties to consider the trade-offs between conflicting objectives.

The study involves biological oceanographic processes and global climate. These themes are concerned with understanding environmental variability which has a major impact on human quality of life. Changes that occur in marine and terrestrial biodiversity, over time and space, will be studied to understand biological responses to environmental variability, as well as the ecosystem level, and to differentiate between the effects of natural and human induced influences on biodiversity. Climate resilience is generally defined as the capacity for a socio-ecological system to absorb stresses and maintain function in the face of external stresses imposed upon it by climate change, and secondly to adapt, reorganize, and evolve into more desirable configurations that improve the sustainability of the system, leaving it better prepared for future climate change impacts.

1.4.1 Penguin Colony Suitable Habitat Site Selection

Climate change plays a huge role in optimal suitable habitat conditions, as will be seen from the modelling, for sea-surface, as well as land temperature. There is an optimal climate range which the penguins prefer.

Penguins generally live on islands and remote continental regions free from land predators. Here, their inability to fly is not detrimental to their survival. These highly specialized marine birds are adapted to living at sea - penguins spend a large amount of their time at sea. Penguins enjoy nutrient-rich, cold water currents that provide an abundant supply of food.

1.5 Existing Approaches for Conservation Planning

There are many management actions taking place, some will now be described.

1.5.1 Provided Areas of Protection

One form of spatial selection is effected through the establishment of Marine Protected Areas (MPAs). Some ecologists advocate for huge sections of the ocean to be designated no-fishing zones (Pauly et al. 2003; Pauly 2009). The hope is that no-take marine reserves protect habitat and biodiversity, buffer against uncertainty in stock assessments, and ultimately increase fisheries yields (Attwood

et al. 1997). However, Agardy et al. (2011) review several reasons why MPAs may not produce the benefits desired. For example, MPAs are unlikely to be of much benefit to fisheries of highly mobile species (Edwards et al. 2008), or to ecosystems when there is little bycatch or habitat impact (Hilborn et al. 2004b). Also, large no-take reserves located near traditional fishing communities may necessitate longer fishing trips, increasing both cost and risk to the humans.

1.5.2 Rehabilitation of Oiled Birds

We should continue to maintain the functions of the Southern African National Foundation for the Conservation of Coastal Birds (SANCCOB) oil spill rehabilitation centre. SANCCOB was formed more than twenty years ago, to rescue penguins and other birds from oil spills and other disasters. It operates a rescue and rehabilitation centre for injured seabirds, near Table View in Cape Town. SANCCOB is funded solely by membership fees and public donations, and has been scientifically proven to be the most successful sea bird rehabilitation centre in the world. In 1994, when the tanker, the Apollo Sea, was wrecked off the Cape Town coast, about 10 000 birds were oiled. About half of these were saved. Much was learnt from this and other disasters. When another major oil slick threatened the penguins after the bulk ore carrier, Treasure, sank off Robben Island in June 2000, an even larger rescue operation was conducted. Over 18 000 oiled penguins were rescued and cleaned. More than 19 000 de-oiled penguins were trucked to Port Elizabeth, where they were released. It was hoped that the oil would have dispersed by the time they returned home. They proved to be efficient navigators.

Another rehabilitative centre where injured, diseased or distressed birds can be treated and rehabilitated, is the African Penguin and Seabird Sanctuary (APSS). It is a Dyer Island Conservation Trust (DICT) project based in Gansbaai, opened in February 2015, aiming to provide local marine avian species with a local rehabilitative centre. APSS has been set up to assist the endangered African penguin colonies of Dyer Island. Here, the species has declined dramatically over 30 years, by almost 90%. The other nearby colony is Stony Point at Betty's Bay. This facility has a fully equipped laboratory and a veterinarian on standby. Thus, we can immediately treat any birds and thereby increase their survival rate.

1.5.3 Active Management Programs

Management to control the population size of predators needs to be investigated (Crawford et al. 2006; David et al. 2003). In the absence of conclusive data, a precautionary approach will be adopted. Otherwise, management interventions that may be adopted, such as culling, removal or relocation of predators, must be used only where sound, relevant scientific data is used as a basis for these decisions (DEA biodiversity management plan, Makhado 2009).

Artificial nests are provided in some colonies (Sherley et al. 2012). The benefits of these are unclear though. The reproductive success of provisioned colonies are similar to that of colonies using natural

burrows or open scrape nests. Provisioned birds did, however, show greater reproductive success than those nesting under vegetative cover.

Design and implementation of actions is used to control the spread of disease within breeding colonies (Crawford et al. 2006). Namibian breeding localities also need to be protected (Ellis et al. 1998). Plans are developing to conserve pelagic fish resources (Harrison et al. 1997), namely through management of the purse-seine fishery (Crawford et al. 2006). There needs to be worked on the prevention of oil spills from the illegal cleaning of ship tanks (Harrison et al. 1997). Work has also been carried out to eliminate feral cats from Bird, Dassen and Robben Islands and implement measures to preclude the introduction of rats to any colonies (Ellis et al. 1998, Crawford et al. 2006).

Investigated reintroduction techniques (Ellis et al. 1998) and established captive breeding populations are used to assist with future reintroduction or supplementation efforts. Assessments are performed to see whether climate change is a factor in the shifting of prey populations (Koenig 2007). Considerations of the idea of establishing no-fishing zones around breeding islands (Koenig 2007, L. Underhill per Koenig 2007), trans-locating birds in reaction to shifts in food availability (L. Underhill per Koenig 2007), and maintaining suitable breeding habitat (Crawford et al. 2006) are being investigated. Work is being performed on establishing and then monitor "trial colonies" close to current concentrations of food resources (R. Wanless in litt. 2010).

Several management actions that have been implemented to conserve the African penguins include formal protection of breeding colonies by converting areas with known breeding sites into nature reserves and national parks, prohibiting the collection of guano and eggs, establishing marine protected areas where fishing is prohibited, conducting ongoing research to monitor population trends in relation to prey availability and disease outbreaks, active management of population sizes of predators, artificial care of abandoned chicks, providing artificial nests and rehabilitating sick birds, and investigating the viability of artificial insemination. Some of the leading organisations in the conservation of the African penguin include the South African Foundation for the Conservation of Coastal Birds (SANCCOB), Dyer Island Conservation Trust (DICT) and South African Marine Rehabilitation and Education Centre (SAMREC).

Raising awareness of the decline of the African penguin, and other environmental matters, is a contributing factor towards conservation. This year I was part of Waddle 2017, in which 16 environment enthusiasts walked from the African Penguin Seabird Sanctuary in Gansbaai to Boulders beach (130km) in aid of this. We waddled past Onrus River, where there is a Marine Protected Area (Haarder Bay) and where penguins are regularly noticed in the sea. The penguins from Dyer Island and Stony Point (Betty's Bay) most probably come to this nearby reserve, where there are still plenty of fish on which they can prey.

1.5.4 Ongoing Investigation and Research

One should monitor population trends at all colonies (Ellis et al. 1998). There is a project being conducted aimed at micro-chipping penguins with transponder devices, to gather data on penguin population survival and movement patterns (SANCCOB). One should also initiate more research into the impacts of fishing and predation (Ellis et al. 1998). Ongoing research should be established to understand the penguin feeding behaviour and prey availability (eg. by Koenig 2007). It is important to assess the impacts of climate change on the population of prey species (Koenig 2007). As the prey shift and climate change occurs, ongoing research is necessary.

The rapid degradation of ocean ecosystems dictates the urgent necessity for spatial conservation planning and management measures. These could be modified later, with the acquisition of new information. A recently emerged approach to conservation planning is the use of Species Distribution Models (SDMs). Mapping habitat suitability for species, using SDMs, has been increasingly applied as a conservation planning tool. This is especially used for predicting the impact of climate change and land use changes on biodiversity. These models can provide insights into systematic conservation planning, for use in decision making processes.

1.6 Research Questions and Objectives

This research explores the possibility of relocating, or establishing new penguin colonies, taking into consideration habitat suitability in human-modified landscapes. To achieve this, SDMs are developed using the species' occurrence information to (1) map habitat suitability of African penguins along the African coastline; (2) identify and test the relative contribution of environmental variables ecologically relevant to the species' habitat suitability, thus contributing to understanding the reasons for the current decline; and (3) use the predicted habitat suitability, incorporating expert opinion, to make suggestions for establishing the new colonies.

In Chapter 2 I focus on the importance and explanation of SDMs and MaxEnt. The demography of the African penguin is investigated in Chapter 3. In Chapter 4, the modelling method is described, along with the species' datasets and relevant environmental rasters. A raster (also called a "grid") is a spatial (geographic) data structure that divides a region into rectangles called "cells" (or "pixels"), that can store one or more values for each of these "cells". Each "cell" or "pixel" represents an area on the Earth's surface. Applying this knowledge, appropriate seasonality maps are developed and shown in Chapter 5.3: Suitability Mapping.

The aim of this study is to better understand the effects of drivers of change on the African penguin colonies. The establishment of a sustainable management plan for the African penguin species colonies, by consolidating different approaches, will be investigated. Predictors of drivers of spatial variability in the conservation of the African penguin, by implementing the associated parameterisations, will be simulated in this work.

Chapter 2

Species Distribution Models (SDMs)

A SDM is a conceptual model of the abiotic (eg. physical barriers, climate, lack of resources) and biotic (eg. competition, predators, parasites) factors controlling species distributions in space, time and scale (Franklin, 2010). It is a predictive map of species distributions, observations of species occurrences with environmental variables thought to influence habitat suitability, and therefore species distribution. It has also been referred to as environmental, bio-climatic, or species niche modelling, and habitat suitability modelling, correlative models and spatial prediction models. SDM is preferred, as it predicts geographic distribution, rather than environmental (niche) space and the true "niche" is never fully specified or confirmed. Species distribution models provide the modelling environment, where important predictor variables are investigated for the species' distribution, and a suitable habitat map is obtained.

Data on species occurrences in geographical space, and digital maps of environmental variables representing those factors thought to control species distributions, is represented. It is a quantitative or rule-based model, linking species occurrence to the environmental predictors. A Geographic Information System (GIS) for applying the model rules to the environmental variable maps, in order to produce a map of predicted species suitable habitat, as well as data and methods for evaluating the error or uncertainty in the predictions, is used.

Predictive distribution maps are also required for many aspects of resource management and conservation planning. These applications include biodiversity assessment, biological reserve design, habitat management and restoration, and species and habitat conservation plans. Also, population viability analysis, environmental risk assessment, invasive species management, community and ecosystem modelling, ecological restoration, invasive species risk assessment and predicting the effects of climate change on species and ecosystems, can be explored.

The expected form of the response functions, data on species occurrence (location) in geographical space (a measure of presence, but this can also be habitat use, abundance, or some other property, or expert knowledge about habitat requirements or preferences) is given. Digital maps of environmental variables representing those factors (or their surrogates) determining habitat quality, or correlated with it, is shown. These are generally derived from remote sensing, from spatial models of environ-

mental processes, or from some other source, and stored in a GIS. SDM in this research, is a model linking habitat requirements or habitat use (species occurrence) to the environmental variables. The model can be statistical, descriptive, logical, or rule-based (Burgman et al., 2005). Tools for applying the model (rules, thresholds, weights, coefficients) to the values of the mapped environmental variables, to produce a new map of the metric of species occurrence is produced as a GIS (Tobler, 1979).

There are two main approaches for predicting species' niches (Gallien et al., 2010). Firstly, there is the bottom up approach (mechanistic), which uses the physiological characteristics of a species to determine their suitable habitat. Implementation of mechanistic species distribution models requires knowledge of how environmental change influences physiological performance. Ecological variability (e.g., biomass, species richness) is often applied to spatial prediction in other domains, for example, predicting the likelihood of deforestation (Ludeke et al., 1990), urban growth, or fire risk. Secondly, there is the top down approach (correlative), which focuses on the species-environment relationship and the associations between the species' distribution and the environmental factors. Climate is often modelled as the main driver behind species' distributions. Their distributions are in actual fact co-determined by climate, physical structures, disturbances, and biotic and abiotic interactions. This thesis looks at the latter approach to SDM. Correlative species distribution modelling is the most commonly applied approach for predicting effects of climate change on biodiversity, which is one of the major factors contributing towards the decline of the African penguin.

Population viability analysis (PVA) often requires spatially explicit information about the distribution of habitat (location, size and quality of suitable habitat patches), and this can be derived using a SDM relevant to the species under consideration (Akcákaya, 2000). PVA can incorporate landscape dynamics (Pulliam et al., 1992; Lindenmayer and Possingham, 1996; Akcákaya and Atwood, 1997; Kindvall et al., 2004), such as changing carrying capacities of habitat patches through time. SDMs may be used, in this case, to provide the initial conditions (spatial distribution of suitable habitat), or to provide maps of suitable habitat as different time steps, whose changes are driven by landscape dynamics resulting from natural disturbance, land use change or climate change (Akcákaya et al., 2004, 2005; Keith et al., 2008). Changes in natural systems which can be attributable to anthropogenic climate change are now well documented (Walther et al., 2002; Root et al., 2003; Parmesan, 2006; Rosenzweig et al., 2008).

The use of multiple models is highly recommended as a method of addressing the interactions between potential habitat shifts, landscape structure (dispersal barriers caused by land use patterns, landscape patterning caused by altered disturbance regimes), and demography for a range of species functional groups. This method is an effective way of developing guidelines for assigning various degrees of threat to certain species (Keith et al., 2008).

It has been suggested that environmental envelope-type models, using presence-only data, tend to depict potential distributions (suitable habitat), and are more suitable for extrapolation, while more complex models that discriminate presence from absence, tend to predict realised distributions (oc-

cupied habitat), and are more suitable for interpolation (Jim enez-Valverde et al., 2008; Hirzel and Le Lay, 2008).

Fundamental (potential) niche areas is used in response to environment in the absence of biotic interactions. Realised (actual) niche takes into consideration environmental dimensions in which species can survive and reproduce, including biotic interactions. Sober and Peterson (2005) argue that SDM based on coarse-scale climate variables (bioclimatic niche modelling) describes the species fundamental niche. This concept is elaborated by Hirzel and Le Lay (2008) who noted that biotic interactions tend to occur at short distances. Also, that dispersal limitations and fine-scale environmental heterogeneity allow inferior competitors to evade negative interactions by persisting in competitor-free locations. Thus, they conclude, the realised and fundamental niche may not differ that much in practice, especially when predicted from coarser-scale environmental factors, such as climate.

If a model of a geographical distribution is conditioned on a continuous ecological variable, such as biomass, species richness or species abundance (for example, Meentemeyer et al., 2001; Cumming et al., 2000b; Thogmartin et al., 2004; Bellis et al., 2008), then that "dependent variable" is the attribute being predicted. The resulting prediction is in units of grams per m^2. Species per km^2 or individuals per km^2 , for example.

Predictors of drivers of spatial and temporal variability in the conservation of the African penguin, with associated parameterisations, will be studied. The aim is set at predicting a biotic variable (e.g. presence) as a function of explanatory variables. The biotic variable is set as the dependent variable and the predictors as independent variables. Yet, several terminologies exist in the scientific literature: response or dependent/criterion variables which is typically continuous of nature/discrete categorical; vs predictor, explanatory, or independent variables, covariates, inputs; e.g., estimates of climate (marine and terrestrial), currents, topography, and soil for plants (vegetation); temperature, salinity and prey abundance for marine fishes. My specific model will be described in Chapter 4: Methods.

A continuous predictor variable is sometimes called a covariate, and a categorical predictor variable is sometimes called a factor (penguin presence). Usually, you create a plot of predictor variables on the x-axis and response variables on the y-axis.

This dichotomy reflects the logics of regression analyses where a response variable is considered "dependent" of explanatory (or independent) variables. The independent variables are considered uninfluenced by the dependent variable, meaning that there is no immediate feedback. Yet, this dichotomy reflects also the biological logics of the regression modelling approach. We attempt to explain, for example, the presence of a species from biotic and abiotic site factors. Therefore, the presence of a species is considered a physiological or mechanistic logic of these site factors, or in other words, a causal function of the explanatory variables based on the niche requirements of a species. The regression itself does not distinguish between correlative and causal relationships. As soon as a variable is significant in a regression, it can be seen as a statistical predictor, even if the "biological explanation" is irrelevant or wrong. Thus, the outcome largely depends on the experimental design

and context, to determine causal or correlative relationships.

Human activity is the dominant cause of the increase in greenhouse gases in the atmosphere over the last 150 years. The largest sources of greenhouse gas emissions linked to human activity, include from farming practices and burning of fossil fuels for electricity, heat and transportation.

Spatial conservation prioritisation addresses the challenge of how we can best allocate our limited conservation resources, in order to maximise their impact. It can be used for different species and changed to adapt future data. Decision-making in conservation should be efficient and effective, as time and resources are typically limited. Conservation planning is one process by which stakeholders collaboratively make decisions, when attempting to ensure the persistence of biodiversity. Spatial prioritization is the activity of applying quantitative data to spatial analysis, to select locations for conservation investment, and it is a distinct process within conservation planning. The use of experts in spatial prioritization, and more generally in conservation planning, is widely accepted and advocated, but there is no general operational model for how best to involve them. Acceptable standards of practice in selecting experts, and in applying specific techniques for eliciting expert knowledge, need to be developed and tested in different contexts to ensure robust and defensible results of spatial prioritization processes. Although experts and expert knowledge have limitations, including them in spatial prioritization can produce many benefits, such as increased robustness of decisions and time and cost savings. Timeous, decisive, cost-efficient and sound decision-making is essential when attempting to stem the continued loss of biodiversity across the world, in South Africa and specifically in relation to the African penguin. The use of SDMs in the decision making process is indicated in Figure 2.1. Although widely used, very little research has been conducted into the role of experts in spatial prioritization processes.

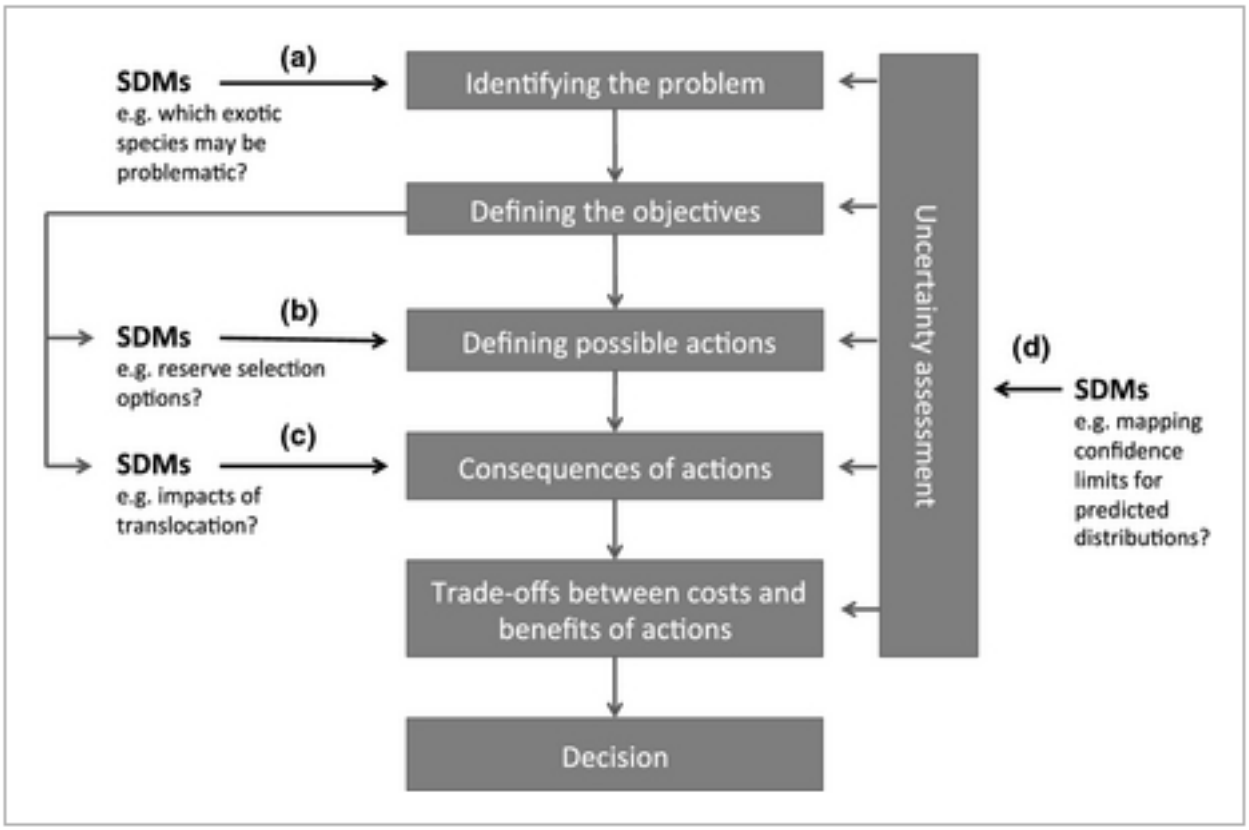

Figure 2.1: A structured decision-making process with indication of potential entry points for the use of SDMs (Source: Gregory et al. 2012).

The most effective and cost efficient approach to integrating spatial prioritization software with expert knowledge will be incorporated. Some modelling methods are discussed in Subsection 2.1.

2.1 SDM Methods

Relevant modelling methods include straightforward environmental matching models such as BIOCLIM and DOMAIN. Also there are Generalized Linear Models (GLM) where the initial regression base is SDMs (Elith and Leathwick, 2009). Other increasingly complex models, incorporating non-linear relationships such as Generalized Additive Models (GAM) and Maximum Entropy models (MaxEnt) are used. Most SDM methods are regression-like. Additive combinations of predictors can model species' abundance. Multivariate Adaptive Regression Splines (MARS) use piecewise linear fits rather than smooth functions. This allows for faster implementation than GAMs (Elith et al., 2006). Some of the initial SDMs use presence-only data (such as BIOCLIM, DOMAIN). As SDMs developed, most methods started to incorporate absence data as well, leading to an improvement in model accuracy. Machine learning and Bayesian methods are the most recent developments. These allow for sophisticated model fitting abilities. The complication is that these processes are more computationally intensive. Machine learning techniques are more complex and often viewed as "black

boxes". This requires greater insight into the ecological application and functioning of said techniques. When studying the response functions (the relationship between species' occurrences and their environment), BRT and MaxEnt fitted the separate functions best according to their Area Under Curve (AUC) performance as seen in Figure 2.2. AUC is discussed in detail in Section 4.6.

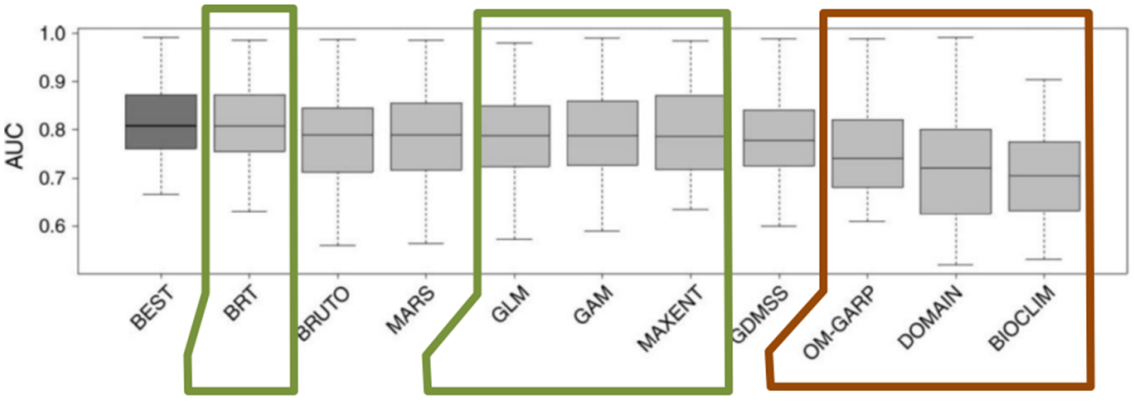

Figure 2.2: SDM Methods (Source: Guisan et al., 2007. Ecological Monographs, 77: 615-630).

According to research, due to its good, comparative performance to other methods (Figure 2.2), for the SDMs, the machine learning technique, Maximum Entropy (MaxEnt) has been used in the R statistical programming language. It is also used because we are working with presence-only data. MaxEnt even performs well for modelling incomplete and biased data, with limited sample sizes which could be common problems in studies.

2.2 Definition of MaxEnt Property of a Distribution

The MaxEnt program, used in my model, is a species distribution modelling (SDM) tool. For my model it is used to predict where a species' colonies can possibly occur, based on the environmental conditions specified. It is applied to known sites of where the colonies presently exist and also extrapolated to other areas of which the suitability is being explored. Modelling with the MaxEnt method creates the maximum entropy (most spread out or as even as possible, as indicated in Figure 2.3), under certain constraints. The constraints are the limitations applied to the possible values of the environmental variables, relating to the distribution of the colonies in this study, those of the species of the African penguin.

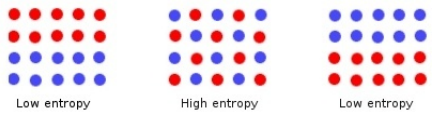

Figure 2.3: The first and last pictures show low entropy, as they have a well ordered, or separated, indication of blue and red objects (variables in my model). The middle one indicates high entropy: it has evenly or uniformly placed red and blue objects. Thus, maximum entropy is achieved when we have an uniform distribution of things or in other words when they have the most evenly spread out distribution. (Source: https://www.quora.com/What-is-maximum-entropy-in-the-simplest-terms).

The Maximum Entropy principle states that when fitting a probability distribution to data, it is subject to a set of constraints. The most likely occurrence maximizes entropy (Jaynes, 1957). This approach identifies the habitat suitability using all available information, the data and constraints, while making the fewest assumptions.

Without adding constraints, the simplest distribution is the one as even as possible / closest to uniform, the most uninformative prediction, and has the highest entropy. In other words, if the only information about a probability distribution available is incomplete, the assumption is that it is a distribution as close as uniform from the available information. MaxEnt is a technique that makes a prediction from the data available. The primary property is that it takes into consideration the known, but nothing about the unknown data and constraints. The aim is to forecast the environmental suitability for the penguin colonies, as a function of the environmental variables chosen, and to represent it in a geographic distribution map. The environmental variables chosen in my model are: sea surface temperature, mean temperature, chlorophyll count, precipitation and precipitation seasonality (coefficient of variation, which is the ratio of the standard deviation to the mean). There are other variables which could be used; however, the ones I used for this study are the ones I deemed to be the most important factors with regards to predicting habitat suitability for African penguin colonies. The known environmental conditions with regard to these chosen variables, at the present penguin colony locations, are then used to extrapolate to indicate the degree of suitability of the habitat of the whole study area (from the Namibian to the P.E. coastline).

The sample points are made up of pixels where we know the penguin colonies exist. The features are made up of environmental (including climatic) variables, and functions thereof. The sample average (random samples drawn independently, with replacement) of the environmental variables at the locations of the penguin colonies, is taken. The MaxEnt method applies constraints, then the model indicates the environmental habitat suitability conditions to the extent within the areas under study.

The location-based model built from the MaxEnt model, starts from a uniform distribution of probability. This is for each cell of the raster. The constraints then determine to what extent it moves away from this distribution. It improves the fit to the model, iteratively, until the gain, which is related to deviance described next, is saturated. The gain is a likelihood statistic which shows how closely the

model occurs to its presence samples, thus maximizing the probability of presence, corresponding to the background data (no species specific information is needed, described in detail in Chapter 4.3.1). The only species data available are the geographical coordinates of sites where the penguin colonies were observed. The space on which the MaxEnt probability distribution is defined, is made up of the pixels of the study area, for presence-only species distribution modelling.

Compared to a uniform distribution the MaxEnt distribution, created by the MaxEnt program, fits the presence data much better (Yost et al. 2008). The MaxEnt distribution is discrete (specific values), partitioning the study region into a set of grid cells and assigning a probability to each cell. The probability associated with a cell is the probability that, given that a colony (in my model's case) of the species was observed, that colony derived from that cell (Phillips et al., 2006). To apply the Maximum Entropy principle, the constraints are defined using the functions of environmental variables (called environmental features), which is further discussed in Chapter 4.4.

The aim of MaxEnt is to approximate a predicted probability distribution, by identifying the probability distribution of maximum entropy, which is the most unconstrained one (Jaynes 1957), limited to the required constraints that represent our partial information about the predicted distribution. The functions of the environmental variables, the features, are used for the predicted distribution. The constraints are required to indicate that the value of each feature should match or be close to its sample average. The sample average is the average value for a set of sample points, taken from the presence sites.

As an example, for the feature "annual temperature", the constraint that goes with it states that the mean annual temperature of the model should be close to the average measured temperature. The set of constraints typically specifies the model incompletely (under-specifies), among all probability distributions satisfying the constraints, thus we choose the one of maximum entropy, i.e. the most unconstrained one (Jaynes 1957). When adding constraints or features it lowers the maximum entropy.

While earlier papers (Philips et al. 2006, Phillips and Dudík, 2008) have outlined MaxEnt as approximating a distribution across geographic space (an occurrence-based statement over an exact area of landscape, a grid of pixels), the paper: "A Statistical Explanation of MaxEnt for Ecologists", by Elith et al. 2010, places emphasis on comparing probability densities in covariate space (independent, environmental variables). The latter paper defines MaxEnt as a model that minimises the relative entropy between two probability densities - the one estimated from the presence data and one estimated from the landscape.

There are effective deterministic algorithms that have been developed that converge to the optimal, maximum entropy probability distribution. The MaxEnt probability distribution has a concise mathematical definition, which I will now define.

2.3 Mathematical Formulation of the MaxEnt Principle

The real distribution of a species is shown as an unknown probability distribution π, which occurs over a set X of sites in the study area, in the maximum entropy density estimation. To explain why these models are known as density estimation, it is taken from the fact that we are estimating the density of observations (presences) across the terrain (Fithian and Hastie 2012). X is the finite set of pixels, or points, in the study area, and indicate the measured occurrence localities for the species, African penguin colonies, as sample points $x_1, ..., x_m$, taken from the unknown probability distribution π. π assigns a non-negative probability $\pi(x)$ to each point x, where these probabilities adds up to one.

We denote $\hat{\pi}$, which is our estimation of π, and is also a probability distribution.

$\hat{\pi}$ has entropy which is outlined as follows:

$$H(\hat{\pi}) = -\sum_{x \epsilon X} \hat{\pi}(x) ln \hat{\pi}(x)$$

The Maxent algorithm maximizes H for a given set of probabilities it seek to calculate. These probabilities are related to the output of the MaxEnt model's environmental suitability scores. Entropy is non-negative. The higher the entropy of the distribution, the more choices involved, in other words the distribution is less constrained. The maximum entropy principle states that the best way is to make certain that the estimation meet the needs of any constraints on the unknown distribution that we know of, and following those constraints, the distribution should have maximum entropy (Jaynes, 1957). It can also be explained that no unfounded constraints should be placed on $\hat{\pi}$.

2.4 Explanation of How the Machine Learning Algorithm Helps to Find the Maximum Entropy Solution

To undertake this, the constraints on the unknown probability distribution π, is stated. To clarify again we have features, the set of known functions of environmental variables $f_1, ..., f_n$ on X. We suppose the details known about π is described by the forecasts (averages) of the features under π. Each feature f_j allocate a real value $f_j(x)$ to each point x in X. The prediction of the feature f_j under π is defined as:

$$\sum_{x \epsilon X} \pi(x) f_j(x)$$

For the prediction of f under p, for any probability distribution p and function f, we use the notation $p[f]$, and it is indicated by $\pi[f_j]$.

The feature predictions $\pi[f_j]$ will be estimated using a set of sample points $x_n, ..., x_z$. These are drawn

independently from X, with replacement, in line with the probability distribution π.

The sample average of f_j is

$$\frac{1}{m}\sum_{i=1}^{m} f_j(x_i)$$

which can also be written as $\tilde{\pi}[f_j]$, where $\tilde{\pi}$ is the uniform distribution on the sample points. It is also used as an estimate of $\pi[f_j]$.

From the maximum entropy principle, we are looking for the probability distribution $\hat{\pi}$ of maximum entropy, following the constraint that each feature f_j has the same mean under $\hat{\pi}$ as detected by the sample, i.e.

$$\hat{\pi}[f_j] = \tilde{\pi}[f_j] \tag{2.4.1}$$

for each feature f_j.

The mathematical theory of convex duality shows (Della Pietra et al., 1997) that this formulation uniquely determines $\hat{\pi}$. In mathematical optimization theory, the duality theory specifies that optimization problems may be viewed from either of two perspectives: the primal or the dual problem. $\hat{\pi}$ has a different characterization, which will be described next.

Let us regard all probability distributions in the form of

$$q_\lambda(x) = \frac{e^{\lambda.f(x)}}{Z_\lambda} \tag{2.4.2}$$

where λ is a vector of n real-valued coefficients or feature weights, f corresponds to the vector of all features, and Z_λ is a normalizing constant that guarantees that q_λ adds to 1. These distributions are recognised as Gibbs distributions.

The feature values at x are the attributes which the value of the MaxEnt model q_λ, at a site x depends on. Therefore, it only depends on the environmental variables at x. The MaxEnt model, which we initially defined in accordance to the set X of training sites, therefore, can also be "projected" to other sites where the same environmental variables are available.

According to convex duality, the MaxEnt probability distribution $\hat{\pi}$ is equivalent to the Gibbs probability distribution q_λ, which maximizes the likelihood or probability of the sample points. It also minimizes the negative log likelihood of the sample points

$$\tilde{\pi}[-ln(q_{\lambda)}] \tag{2.4.3}$$

which is equivalent to

$$lnZ_\lambda - \frac{1}{m}\sum_{i=1}^{m}\lambda f(x_i)$$

and is termed the "log loss".

The MaxEnt model can likely cause overfitting of the training data. The issue is based on the fact that the sample feature means will normally only approximate the true means, not equal them. The means under $\hat{\pi}$ should therefore be confined to only be close to their sample values.

The means under $\hat{\pi}$ should be limited to be close to their sample values. This can be done by relaxing the constraint in equation (2.4.1), by substituting it with

$$|\hat{\pi}[f_j] - \tilde{\pi}[f_j]| \leqslant \beta_j,$$

for each feature f_j and for some constants β_j.

The dual characterization is now changed in a form of l_1-regularization. The MaxEnt distribution is now shown as a Gibbs distribution that minimises

$$\tilde{\pi}[-ln(q_\lambda)] + \sum_j \beta_j|\lambda_j|.$$

The first term is expressed as the log loss as in equation (2.4.3). The second term penalizes the use of bigger values for the weights λ_j. Because of regularization, the MaxEnt model is forced to concentrate on the most important features. l_1-regularization usually create models with few non-zero λ_j values (Williams, 1995). These models usually won't cause overfitting, as they have fewer parameters.

I will now clarify how the Machine Learning algorithm is used in MaxEnt. The MaxEnt probability distribution starts from the uniform probability distribution. This is where $\lambda = (0, ..., 0)$. It then frequently make modifications to one or more of the of the weights λ_j by means that the regularized log loss decreases. Regularized log loss is a convex function of the weights, so no local minima exist. Various convex optimization methods are being used for modifying the weights in a way that assure convergence to the global minimum. This establishes the maximum likelihood formulation for finding the MaxEnt probability distribution, we derive the likelihood with regard to each feature weight. Refer to Section 2.3 for the algorithm used for the MaxEnt modelling approach.

We use this unconditional maximum entropy models as described, as both presence and absence data would be needed to train a conditional model of a species' distribution.

2.5 How the MaxEnt Model is Used

MaxEnt is used in the following manner (Rivera et. al. 2017):

(1) a likelihood function P (positive value) shows the presence of the African penguin colonies, on a set x of points in the study zone, where P(x) sums to one.

(2) the model of P has a number of constraints, acquired from the sample data of presence locations.

(3) the restrictions are indicated as a simple function of known environmental variables, *f*(variable).

(4) the average forces of each function, of each variable, are close to the actual average of the variable at the presence locations in the MaxEnt method.

(5) of the viable options available, a particular combination of features is chosen to maximize the entropy function. The entropy function do not take into consideration restrictions that do not provide the model with relevance. This allows for optimal selection of variables and functions depending on their significance.

2.6 Relationship Between MaxEnt and Other Modelling Approaches

The MaxEnt model uses the same concept of logistic regression (a probabilistic model for binomial cases). Multinomial logistic regression is a method that uses logistic regression for multiclass problems - problems with more than two possible discrete outcomes. When we explore the constrained optimization problem of MaxEnt, it gives us the log-linear form used in multinomial logistic regression. By using a multinomial model, the output is in the form of relative occurrence rate (RORs), and automatically sums to one, covering the terrain. Density estimation is used here.

There are methods that are similar to the MaxEnt method for modelling species distributions. In specific, generalized linear models (GLMs), generalized additive models (GAMs) and some machine learning methods. Some of these machine learning methods include Bayesian approaches and neural networks. It is appropriate to compare these broad classes of techniques in the manner they have been applied to presence-only modelling of species distributions, as used in MaxEnt modelling. The MaxEnt model is theoretically most similar to GLMs and GAMs.

A GLM often used, is the Guassian logit model or logistic regression.

The logit of the probability of presence that is forecast is

$$\alpha + \beta_1 f_1(x) + \gamma_1 f_1(x)^2 + \ldots\ldots\ldots + \beta_n f_n(x) + \gamma_n f_n(x)^2 \tag{2.6.1}$$

Here, f_j are the environmental variables, and α, β_j and γ_j are fitted coefficients. Creating product variables is a general method for modelling interactions between variables in a GLM. This is analo-

gous to the use of product features in MaxEnt.

The logit function is stated as

$$logit(p) = \frac{p}{1-p}$$

Using GAM to model probability of occurrence using a logit link function, the logit of the predicted probability can be shown as

$$g_1(f_1(x)) + ...g_n(f_n(x))$$

f_i are environmental variables, and g_i are smooth functions fit by the model, with the amount of smoothing regulated by a width parameter. This form, is similar to the log probability of the pixel x in a MaxEnt model, with threshold features. The regularization has an equivalent effect to smoothing on the otherwise random functions g_i. The shape of the response curve to each environmental variable is decided by the data, for both the MaxEnt model and using GAM.

Not only are there many similarities, but some differences also exist between GLM/GAMs and MaxEnt. The outcome will cause them to make different forecasts. Absence data is required when GLM/GAMs are used to model the probability of occurrence. When presence-only data is used, background pixels must be used instead of true absences (Ferrier and Watson, 1996; Ferrier et al., 2002). The outcome is a relative index of environmental suitability. MaxEnt models a probability distribution over the pixels in the study region, where the pixels without species records should not be interpreted as absences. MaxEnt is also a generative approach, whereas GLM/GAMs are discriminative. Generative methods may give better predictions when the amount of training data is small (Ng and Jordan, 2001).

MaxEnt is similar to other machine learning (systems that are provided the ability to automatically learn and improve from experience without being explicitly programmed) methods in the way it approaches probabilistic reasoning. Regularization, where the use of large values of model parameters reduces the accuracy, can be illustrated by the use of a Bayesian prior (Williams, 1995). MaxEnt is however, rather different from the particular Bayesian species modelling approach of, for example, Aspinall (1992). This machine learning method takes independence of environmental variables, and the assumption is often not met for environmental data.

Environmental Niche Factor Analysis (ENFA) requires similar data as that of the MaxEnt model. The ENFA technique also uses presence data together with environmental data for the entire study area. Both methods, could however use only a random sample of background pixels to improve the running time.

2.7 Comparability of the MaxEnt Method to the Bayes' Theorem

Jaynes stated Bayes' theorem was an approach to calculate a probability, while maximum entropy was a way to assign a prior probability distribution. Maximum entropy density estimation can be compared to robust Bayes estimation. This is from a decision theoretic perspective. In this case the objective of the modeller is to optimize the expected log likelihood. Provided that the only actual attribute known about the true distribution π is that it satisfies a certain set of constraints, obtained from the penguin colonies' locality data, then the strategy which assures the best performance regardless of π, also called the minimax strategy, is to choose the maximum entropy distribution subject to the given constraints (Topsoe 1979, Grünwald 2000, Gruünwald and Dawid 2004).

To explain how π represents the realized distribution of the species, we consider the following (idealized) sampling strategy. An observer picks a random site x from the set X of sites in the study area. The observer then records 1 if the species is present at x, and 0 if it is absent. If we take the response variable (presence or absence) as y, then $\pi(x)$ is the conditional probability $P(x|y = 1)$, i.e. the probability of the observer being at x, given that the species is present.

According to Bayes' rule,

$$P(y = 1|x) = \frac{P(x|y = 1)P(y = 1)}{P(x)} = \pi(x)P(y = 1)|X| \tag{2.7.1}$$

since according to our sampling strategy $P(x) = 1/|X|$ for all x. The overall prevalence of the species in the study area is indicated by $P(y = 1)$. The quantity $P(y = 1|x)$ is the probability that the species is present at the site x. This is 0 or 1 for plants, but may be between 0 and 1 for vagile organisms. For my model, I indicate 1 for penguin colony presence and 0 for colony absence.

Equation (2.7.1) shows that π is proportional to probability of presence. However, if we have only occurrence data, we cannot determine the species' prevalence (Phillips et al. 2006, Ward et al. 2007). We therefore estimate the distribution π, rather than estimating $P(y = 1|x)$ directly. We point out that here x is a site, and not a vector of environmental conditions. This approach differs from more traditional statistical methods, such as logistic regression.

The MaxEnt distribution belongs to the family of Gibbs distributions (Dudík et al. (2004)) derived from the set of features $f_1, ..., f_n$. Gibbs distributions are exponential distributions parameterized by a vector of feature weights

$$\lambda = (\lambda_1, ..., \lambda_n)$$

described before in equation (2.4.2).

2.8 Advantages and Disadvantages of using the MaxEnt Modelling Technique

The MaxEnt technique has many advantages, but only a few drawbacks. The advantages include the following:

(1) Only presence data is needed, in combination with environmental data for the entire study area.

(2) It can put both continuous and categorical data to use, and can include interactions between different variables.

(3) Efficient deterministic algorithms have been developed that can assure to converge to the optimal (maximum entropy) probability distribution, as indicated in Chapter 2.4.

(4) The MaxEnt probability distribution has a concise mathematical definition as defined in Chapter 2.3.

(5) Overfitting can be avoided by using l_1 - regularization, discussed in Chapter 2.4.

(6) The dependence of the MaxEnt probability distribution on the distribution of occurrence localities is clear. Therefore, in future the issue of sampling bias could be addressed formally, as in Zadrozny (2004).

(7) The output is continuous, allowing fine distinctions to be made between the modelled suitability of different areas. This is very valuable for conservation planning. Great flexibility in the choice of threshold is allowed, when binary predictions are wanted.

(8) MaxEnt modelling could also be applied to species presence/absence data by using a conditional model (as in Berger et al., 1996), instead of just using the unconditional model used here.

(9) It is a generative approach, rather than discriminative, which can be an intrinsic advantage when the amount of training data is limited.

(10) MaxEnt modelling is an active area of research in statistics and machine learning. Progress in the field can currently be applied here.

(11) MaxEnt is a general-purpose and flexible statistical method, it can be used for many applications, at all scales.

Some drawbacks of the method includes the following:

(1) It is not as mature a statistical method as GLM or GAM. There are thus fewer guidelines for its use in general, and fewer methods for estimating the amount of error in a prediction. Our use of an "unconditional" model is scarce in machine learning.

(2) The amount of regularization requires further study (e.g., see Phillips et al., 2004), as well as its

effectiveness in avoiding overfitting compared with other variable selection methods (see for example Guisan et al., 2002).

(3) It uses an exponential model for probabilities, which is not inherently bounded above - thus it can give very large predicted values for environmental conditions outside the range present in the study area. Special caution is therefore necessary when extrapolating to another study area, or to future or past climatic conditions. Feature values outside the range of values in the study area, should be "clamped", or reset to the appropriate upper or lower bound.

(4) Special-purpose software, such as maxent.jar (which is put in the java folder of the dismo package), is required, because MaxEnt is not available in standard statistical packages.

2.9 Environmental Variables and Feature Classes in MaxEnt

There are two types of environmental variables that the features in MaxEnt are derived from, these are: continuous and categorical. Continuous variables take random real values, which correspond to measured quantities such as annual precipitation, and minimum temperature. Categorical variables take only a limited number of discrete values such as soil type or vegetation type.

The MaxEnt program (Phillips et al. 2005) uses features of six classes: linear (L), quadratic (Q), product (P), threshold (T), hinge (H), and category indicator features. Hinge features are initiated in the paper: Phillips and Dudík 2008, while the other five classes were introduced in Phillips et al.(2006). Linear, quadratic, product, threshold, and hinge features are acquired from continuous variables.

This list of features are defined next. Our ecological assumptions are the constraints imposed by the features. We are asserting that they represent all the environmental factors that constrain the geographical distribution of the species.

(1) A continuous environmental variable f is a "linear feature". It adds the constraint on $\hat{\pi}$ that the mean of the environmental variable,

$$\hat{\pi}[f],$$

should be close to its observed value, that is, its mean on the sample localities.

(2) The square of a continuous variable f is a "quadratic feature". When it is used with the corresponding linear feature, it exert control over the constraint on $\hat{\pi}$ that the variance of the environmental variable should be close to its observed value, since the variance is equal to

$$\hat{\pi}[f^2] - \hat{\pi}[f]^2.$$

It takes into consideration the species' tolerance for variation from its optimal conditions.

(3) The product of two continuous environmental variables f and g is a "product feature". In combination with the linear features for f and g, it requires from the constraint, that the covariance of those two variables should be close to its observed value, since the covariance is

$$\hat{\pi}[fg] - \hat{\pi}[f]\hat{\pi}[g].$$

Product features therefore uses interactions between predictor variables.

(4) For a continuous environmental variable f, a "threshold feature" is equal to 1 when f is above a given threshold, and 0 otherwise. The proportion of π that has values for f above the threshold should be close to the observed proportion. This is a given constraint. All possible threshold features for f together allow MaxEnt to model a random response curve of the species to f, as any smooth function can be approximated by a linear combination of threshold functions.

(5) For a categorical environmental variable that takes on values v_1.....v_k, we use k "binary features", where the i-th feature is 1 wherever the variable equals v_i, and 0 otherwise. As with threshold features, these binary features constrain the proportion of $\hat{\pi}$ in each category to be close to the observed part. Category indicator features are derived from categorical variables.

Each feature constrains the means, variances, and covariances of the respective variables, in order to match their sample values (Phillips et al. 2006). Extra feature types could be derived from the same environmental variables, thus the list of features is not exhaustive.

2.10 MaxEnt Output Formats

The output represented by the MaxEnt model, is the exponential function $q_\lambda(x)$ (defined in equation 0.2). During model training, it assigns a probability (referred to as a "raw" value) to each site. Raw values are not instinctive, it is in fact hard to interpret "projected" values obtained by applying q_λ to environmental conditions at sites not used during model training. Raw values are also scale-dependent. In the instance where more background data is used, it results in smaller raw values, because they must sum to one over a larger number of background points. For these reasons, raw values have generally been converted into the "cumulative" format (Phillips et al. 2006).

The cumulative format is defined in terms of omission rates predicted by the MaxEnt distribution q_λ. In particular, we consider 0-1 prediction rules that threshold raw output at a level p. Each raw threshold p is changed into the omission percentage c(p) predicted by q_λ for the corresponding rule, i.e.

c(p)=100

$$\sum_{x:q_\lambda(x)\leq p} q_\lambda(x).$$

Thus, if we make a 0-1 prediction from the MaxEnt distribution q_λ using a cumulative threshold of c, the omission rate will be c% for test sites drawn from q_λ. The cumulative format is scale-independent. It is more easily interpreted when projected, but it is not definitely proportional to probability of presence.

Chapter 3

Demography of the African Penguin

Demography is defined as the study of how the population sizes of species change over time and space. The development of population ecology / dynamics comes mainly from this. It includes statistics such as births, deaths, or the incidence of disease, which illustrate the changing structure of populations. The factors affecting the rate of spread and spatial distribution of the penguin population will be investigated. Identification of ecological determinants and mechanisms of the spatial-temporal dynamics for the modelling is also of importance. Biotic interactions, environmental heterogeneity and stochasticity, used to determine how to best design the SDMs, will be looked into.

Reliable estimates of survival and dispersal are crucial to understanding population dynamics. For seabirds / penguins, however, in which some individuals spend years away from land, mortality and emigration are often confounded. Multistate mark-recapture methods reduce bias, as they incorporate movement into the process of estimating survival (study done by Sherley et al. 2014).

Under favourable conditions, juvenile survival can be similar to adult survival in penguins (Saraux et al. 2011b, Dehnhard et al. 2014). Juveniles may struggle in poor conditions because of lower foraging efficiency (Wilson 1985). Their survival rates can show high temporal variability, with important implications for population numbers (e.g. Votier et al. 2008, Dehnhard et al. 2014). Survival of African penguins was usually higher in years of high, rather than low, abundance of anchovy (Whittington 2002). Factors affecting the survival will be discussed in this Chapter.

3.1 Foraging

African penguins feed solitarily or in small to large groups, up to 150 birds (Rand 1960; Wilson and Wilson 1990; Ryan et al. in review). They usually forage at depths < 80m, but may dive to 130m. Dives last on average 1 to 2 minutes. They can hunt co-operatively, swimming rapidly around a school of fish, to compress it (Wilson 1985b; Wilson and Wilson 1990; Ryan et al. in review). Most food is caught between 10h00 and 18h00, with a break in feeding activity around midday (Wilson and Wilson 1995; Petersen et al. 2006; Ludynia 2007; Waller 2011). Birds generally do not feed at night (Wilson 1985a). When breeding, most foraging trips last < 24h. African penguins mainly have short foraging ranges (10-50 km; Wilson 1985). They can perform between 200 and 400 dives in a foraging

trip (Ryan et al. 2007). Foraging effort increases with the growing chicks, and parents brooding large chicks can forage for 3 to 5 days (Ludynia, Waller unpublished data). Outside of the breeding season, birds may travel up to 120 to 350km (Ludynia 2007; Waller 2011), but may even have to travel further for prey (Birdlife).

African penguins forage in dynamic coastal environments. Ocean physical processes have been used to increase the probability of locating penguins' small pelagic prey. Modelling the sea-surface thermal habitat preferences and the dive behaviour of the penguins, in relation to thermoclines, is of importance and has been investigated, such as: "Foraging ecology of the African Penguin (*Spheniscus demersus*) in relation to ocean physical processes" (Rowen Brandon van Eeden). Penguins mainly commute towards cool, nutrient rich waters from a periodic up-welling cell. Also, the shift of small pelagic schools of sardines (*Sardinops ocellatus Pappe*) and Cape anchovy (*Engraulis capensis Gilchrist*), from the West to East (e.g., Crawford et al. 1995) plays a role. The bulk of the fish stock of penguins is now away from the vast majority of fishing capacity and efforts in the Western Cape area of South Africa, where fish processing facilities, and most penguin breeding colonies also, are located. Theory indicates that heavy fishing effort in the West, is leading to less fish in the water around breeding islands to sustain the significant food requirements for breeding birds (Whittington et al. 2005, Pichegru et al. 2012).

Penguins depart in the early morning, before dawn, travelling towards areas of cool, nutrient rich waters. Here, they maximize the time they forage during the day in cooler waters with a higher probability of containing prey patches. Penguins used a correlated random search strategy during foraging, suggesting that they searched continuously for prey. It is likely that penguins are limited by the patchy distribution of prey rather than an abiotic heterogeneous marine environment.

Penguins show flexibility in their foraging behaviour by adjusting their dive behaviour to subsurface thermal structures. Penguins also demonstrated foraging optimization by using temperature cues and behavioural switching, to maximize the probability of locating prey patches on a fine temporal and spatial scale.

Also, when diving, penguins utilized thermoclines that fronted cool waters, as a potential cue to prey. However, their dive depths may also reflect the distribution of their prey. This may aggregate around thermoclines due to increased productivity. Penguins dived deeper, foraging below the thermocline, when the thermocline depth increased. They also responded differently in their dive behaviour under different thermocline structures. For instance, when thermoclines were a diffuse barrier to nutrients and less likely to concentrate prey, birds dived deeper towards the benthos. Warm water intrusions into Algoa Bay, which is from the Agulhas current, resulted in birds diving deeper, in search of cooler, nutrient rich bottom waters.

3.2 Breeding

The African penguin is monogamous (Randall 1983, Crawford et al. 1995a), however, there are exceptions, for instance Deé from uShaka Sea World (Die Burger, 7 July 2017). This could, however, be due to unnatural circumstances, as she was not out in the wild. Penguins breed in colonies, and pairs return to the same site each year. The African penguin has an extended breeding season. Nesting usually peaks from March to May in South Africa, and November to December in Namibia.

African penguins usually breed for the first time at between four and six years of age (Whittington et al. 2005). Once they have bred, adults generally show strong fidelity to colonies and mates, as well as some nest-site fidelity (e.g. Randall et al. 1987; La Cock et al. 1987; La Cock and Cooper 1988; Whittington et al. 2005). First-time breeders have flexibility to emigrate and hence to take advantage of long-term changes in the distribution of food (Crawford 1998). The clutch is usually 2 eggs (sometimes 1, but rarely 3; Crawford et aL 1999, 2000b). Eggs are rounded oval, white and become stained as incubation proceeds. The laying interval is 3 to 3.2 days (Williams 1981, Williams and Cooper 1984). Lost clutches may be replaced and successful breeders may lay again (Randall and Randall 1981, La Cock and Cooper 1988). Incubation starts with the first-laid egg, lasts 38 to 41 days, and is shared equally by both sexes (Rand 1960, Williams and Cooper 1984, Randall 1989). Chicks generally hatch asynchronously, usually about two days apart (Williams and Cooper 1984; Seddon and van Heezik 1991). Chicks are closely attended by adults until about 26 to 30 days when they are mostly left unguarded. They may form crèches of up to 25 chicks (Seddon and Van Heezik 1993, Erasmus and Smith 1974). Chicks fledge when between 55 and 130 days old (Seddon and Van Heezik 1993, Kemper 2006). Often both chicks will fledge from two chick broods, but survival from hatching to fledging is variable and influenced by a multitude of factors such as burrow collapse, exposure, drowning and accidental death in nest and predation by Kelp Gulls *Larus dominicanus*, starvation or heat stress (Seddon and Van Heezik 1991; Barham et al. 2007; Kemper et al. 2007b; Sherley 2010).

About 95% of seabirds are colonial breeders, and become central place foragers (Orians and Pearson 1979) in breeding seasons, in order to brood and feed chicks. Being highly adapted to the environment in which they live, they are sensitive to ecosystem changes (Croxall 1992). Seabirds are thus highly vulnerable to threats at and around their breeding colonies.

3.3 Moult

Moult in seabirds is considered unexpectedly energetically expensive (Hoye and Buttemer 2011). Moult in penguins is unique, since they replace all their feathers in a relatively short period of time compared to that of other birds, ranging from 13 to 40 days depending on the species (Stonehouse 1967). Moult in penguins is an essential feature to them being able to remain waterproof and thus insulated in cold waters while foraging (Stonehouse 1967, Payne 1972). Penguins become hyperphagic during the pre-moult period (Otsuka et al. 2000). The acquisition of sufficient body reserves during pre-moult foraging can be considered a greater priority than at any other time in the annual

cycle (Croxall and Davis 1999, Wolfaardt et al. 2008b, 2009b). Penguins are thus dependant on high and predictable food availability during the pre-moult fattening and post-moult recovery phases. An understanding of timing of moult, including when and where pre-moult fattening and post-moult recovery takes place, is of critical importance for penguin conservation management. Studies of moult patterns in terms of synchrony and seasonality, have shown colony specific variability (Underhill and Crawford 1999, Crawford et al. 2006b, Kemper 2006, Wolfaardt et al. 2009a), which may be attributed to variation in available food resources around the colonies. Ensuring adequate food supply during the pre-moult fattening and post moult conditioning, is essential in order for African penguins to survive the moult (Wolfaardt et al. 2008b, 2009b; Waller 2011).

3.4 Prey

One of the most important current threats to African penguins is considered to be the abundance and availability of prey (Crawford et al. 2007; Crawford et al. 2011b). In the Benguela Upwelling Ecosystem, changes in the relative abundance of sardine and anchovy have been linked to changes in diet, breeding population size and breeding success of various seabird populations, including the African penguin population (Crawford and Dyer 1995; Crawford 2003; Crawford et al. 2006a; 2007; Underhill et al. 2006). The reported eastward shift past Cape Agulhas, in the relative distributions of both sardines (Coetzee et al. 2008) and anchovy adults (Roy et al. 2007,) is considered to have resulted in a mismatch between fish availability and seabird breeding colonies during the summer spawning period, with significant implications for seabirds of the region (Crawford et al., 2011).

Penguins feed mainly on active, free-swimming prey, usually schooling pelagic fish, which they may locate using their olfactory sense (Write et al. 2011). Their diet differs from region to region, but their primary diet includes small pelagic fish such as pilchards, sardines, anchovies (e.g. Engraulis capensis), horse mackerel (*Trachurus capensis*) and round herring. African penguins may supplement their diet with marine invertebrates such as squid, cephalopods (e.g. Randall and Randall 1986), juvenile hake Merluccius sp. (MFMR unpubl. data) and small crustaceans in areas where there is a decrease in fish abundance because of commercial fishing and other factors. In Namibia, African penguins turned to jellyfish and pelagic goby (*Sufflogobius bibarbatus*) after the collapse of small pelagic stocks due to overfishing in the 1970s. A penguin may consume between 540 grams to 1kg (when raising older chicks) of prey every day.

Due to the collapse of a commercial pilchard fishery in 1960, African penguin diet has shifted towards anchovies to some extent. Available pilchard biomass is still a notable determinant of penguin population development and breeding success. While a diet of anchovy appears to be generally sufficient, it is not ideal due to lower concentrations of fat and protein. The interaction of diet choice and breeding success helps the penguins maintain their population size. Penguin diet changes throughout the year. Although parent penguins are protective of their hatchlings, they will not incur nutritional deficits themselves if prey is scarce and hunting requires greater time or energy commitment. This may lead to higher rates of brood loss under poor food conditions. African penguin dietary studies

have relied on the identification of prey remains from stomach contents. Despite all the advantages of this method, it has well known biases. A study has been conducted to assess the African penguin's diet, using stable isotopes, at two colonies in Algoa Bay (south-east coast of South Africa) (Connan et al. 2016). These represent about 44% of South Africa's penguin population. Various samples (blood, feathers, egg membranes) were collected for carbon and nitrogen stable isotope analyses. Results indicate that the trophic ecology of African penguins is influenced by colony, season and age class, but not adult sex. Using Bayesian mixing models it was for the first time shown that adults target chokka squid (*Loligo reynaudii*) for self-provisioning during particular stages of their annual cycle, while concurrently feeding their chicks primarily with small pelagic fish. Carbon and Nitrogen values of the five marine species, sardine, anchovy, red-eye round herring, chub mackerel and squid, were significantly different from each other. On the west coast of South Africa, small pelagics have consistently been recovered in the stomach contents since the end of the 1980s. In the Eastern Cape, small pelagics and squid were identified as the main prey species in varying proportions depending on months and years during the early 1980s. More recently, small pelagics have dominated the stomach contents at both St. Croix and Bird Islands which could be related to the eastward displacement of sardines and anchovy since the mid-1990s.

3.5 Dispersal

Inter-colony movement is increasingly recognized as a fundamental parameter in population dynamics. It may change with time and space depending on environmental and individual conditions (Lewison et al. 2012, Breton et al. 2014, and references therein). Adult African penguins may moult at other colonies (Whittington et al. 2005b), particularly when feeding conditions dictate that they forage far from their breeding colony (Waller 2011, Harding 2013). Some breeding dispersal may occur in seabirds in response to altered environmental conditions (e.g. Distiller et al. 2012). These mechanisms probably account for the movement of adults between places elsewhere in the Western Cape, as well as Robben and Dassen Islands from 1994 to 2003. These colonies were apparently attractive to pre-breeders, prior to 2000, as the sardine and anchovy biomass increased off southwest South Africa (Crawford et al. 2001).

African penguins may be in adult plumage for 4 years before they breed (Whittington et al. 2005a). Birds banded in adult plumage, but not yet breeding, may move from their natal locality to places where feeding conditions are more favourable, subsequently choosing to breed at these non-natal sites (Crawford et al. 1999, 2001). The apparent movement to Robben Island of juvenile and immature African penguins during 1994 to 2003, also conforms with records of chicks banded at Dassen and Dyer Islands later breeding at Robben Island in the 1990s (Whittington et al. 2005c). The population increase here between 1983 and 2000, was primarily driven by immigration linked to the recovery of sardine off South Africa (Crawford et al. 1999, 2001). This natal dispersal appears to offer African penguins limited flexibility to move to localities where current conditions are favourable for breeding (Crawford et al. 1999, in press). However, not all movement of first-time breeders relates to changes in prey availability (Whittington et al. 2005c), and may be density-dependent in seabirds (Crawford

et al. 2007, Gauthier et al. 2010). Thus, the apparent movement of juvenile and immature birds away from Dassen and Robben Islands during 1994 to 2003, and the apparently high natal fidelity during 2004 to 2012, may reflect the marked changes in population size and density at these colonies during the study period (Crawford et al. 2007, Sherley et al. 2014). This may also explain the unexpectedly low movement from Robben and Dassen Islands to other places in the Western Cape after 2003. The researchers anticipated substantial movement away from the west coast colonies in this period, given the changes in sardine availability and the apparent increase in adults moulting at Stony Point after 2001 (Waller 2011), but this was not the case.

3.6 Environmental Variables Incorporating Seasonality

The African penguin is adapted to exploit and cope with a set of factors (including environmental variables), which together determine where on Earth it can live or limit the distribution of the species. Through this study I decided on 5 environmental variables: sea surface temperature, land temperature (penguins occur both on land and in sea), chlorophyll count as a proxy for fish abundance, percipitation seasonality (coefficient of variation), and percipitation of the season under investigation.

I have incorporated seasonality into my study: using annual, summer and winter factors.

Chapter 4

Methods

As discussed in Chapter 3, there are many factors affecting the African penguin species. These dispersal, disturbance and resource factors are depicted in Figure 4.1. For the species distribution models, limiting climatic factors are used.

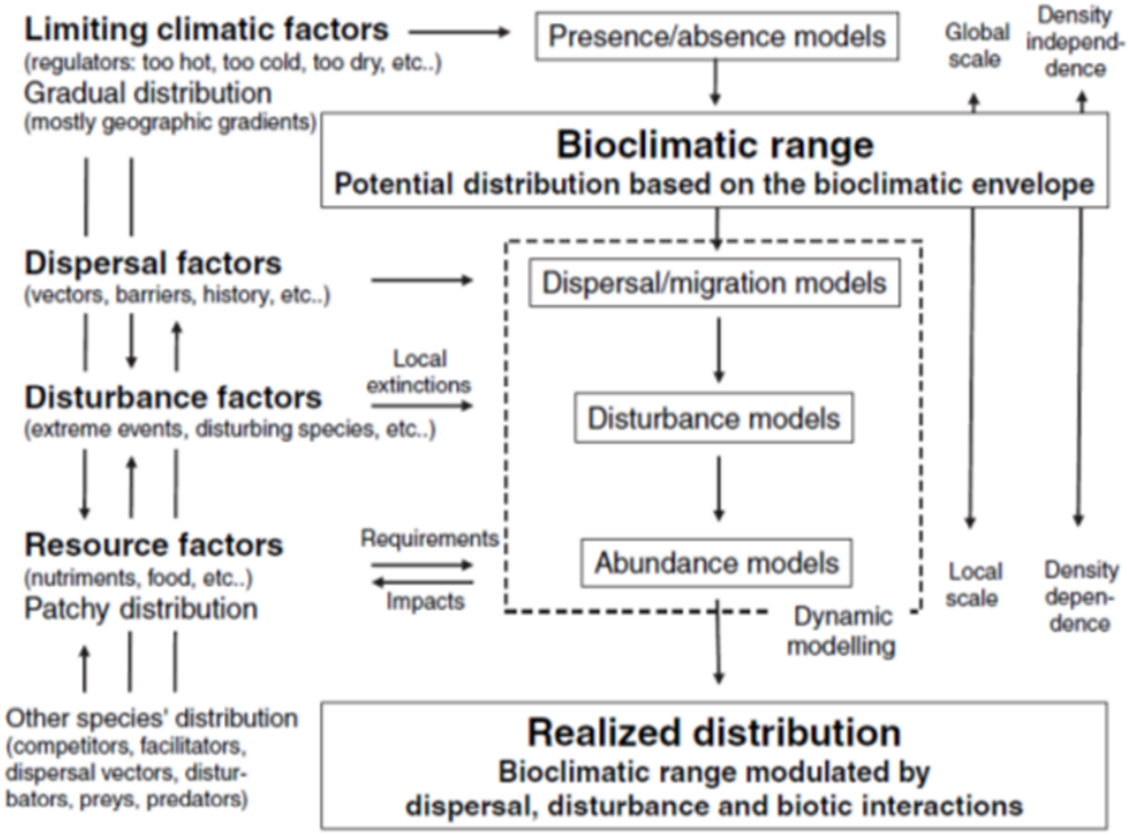

Figure 4.1: General factors affecting species' distributions (Source: Guisan and Thuiller, 2005)

4.1 Implementing Species Distribution Models

Figure 4.2 shows the key steps in implementing the species distribution models.

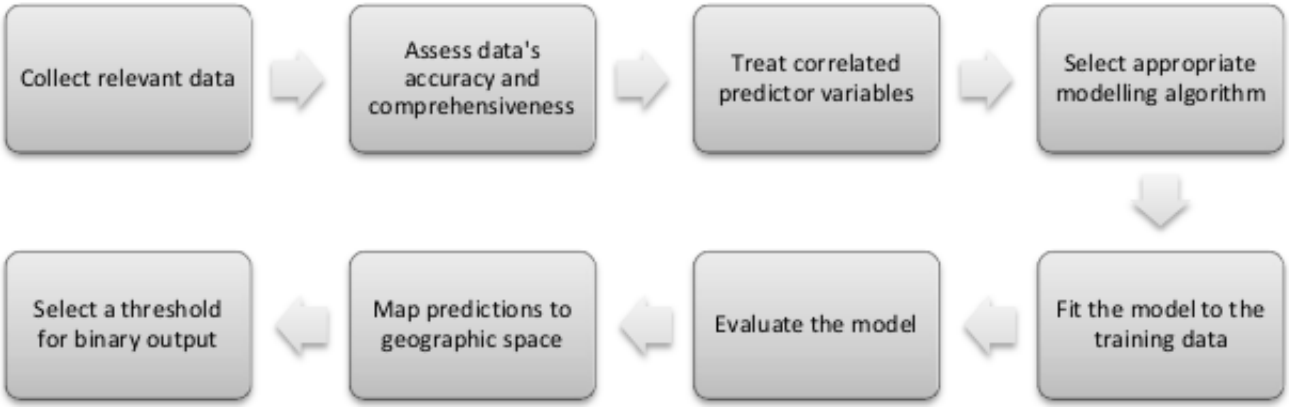

Figure 4.2: Schematic diagram of the key steps in implementing a species' distribution model (Elith and Leuthwick, 2009).

This includes collecting relevant data, assessing the data's accuracy and comprehensiveness, treating the correlated predictor variables, selecting an appropriate modelling algorithm, fitting the model to the training data, evaluating the model, mapping predictions to geographic space, and selecting a threshold for binary output (Elith and Leathwick, 2009). This will be discussed in this Chapter.

Model significance is influenced by the modelling method (Chapter 2.1), the selection of predictor variables, the scale, as well as the extent of extrapolation, which will all be discussed.

4.2 Packages Required for the Code

R statistical programming language has been used for the code. It has a data analysis and graphics environment. We first load the necessary R libraries in the code. Each package (of "R functions") will now be explained (from https://cran.r-project.org/web/packages/):

```
\noindent{library(rasterVis)}
```

This library implements enhanced visualization methods for quantitative data and categorical data, both for univariate and multivariate rasters. It also provides methods to display spatio-temporal rasters, and vector fields.

```
\noindent{library(biogeo)}
```

The above library is used in species distribution modelling as functions for error detection and correction in point data quality datasets. It also includes functions for parsing and converting coordinates into decimal degrees from various formats.

```
\noindent{library(dismo)}
```

Functions for species distribution modelling, that is, predicting entire geographic distributions from occurrences at a number of sites. Also, from the environment at these sites.

```
\noindent{library(raster)}
```

This package implements basic and high-level functions of reading, writing, manipulating, analyzing and modelling of gridded spatial data. The processing of very large files is supported.

```
\noindent{library(spatial.tools)}
```

Spatial functions meant to enhance the core functionality of the package "raster". It includes a parallel processing engine for use with rasters.

```
\noindent{library(sp)}
```

This package is required for classes and methods for spatial data. The classes document where the spatial location information resides, for 2D or 3D data. Utility functions are provided, e.g. for plotting data as spatial selection, maps, as well as methods for retrieving coordinates, for subsetting, print, summary, etc.

```
\noindent{library(corrplot)}
```

A graphical display of a correlation matrix or general matrix. It also contains some algorithms to do matrix reordering.

```
\noindent{library(rgdal)}
```

This incorporates bindings for the Geospatial Data Abstraction Library. The GDAL and PROJ.4 libraries are external to the package, and, when installing the package from source, must be correctly installed first. Both GDAL raster and OGR vector map data can be imported into R. Both data can also be exported. Use is made of classes defined in the sp package.

4.3 Penguin Occurrence Data

We now introduce the code, step by step. The complete R code is attached as Appendix A.

First, we set the working directory and read in the presence data:

```
pres = read.csv("Penguin localities_Frieda.csv")
```

The response / dependent / criterion variable is penguin colony presence (1) or absence (0). Chapter 2.1: SDM Methods, describe how MaxEnt uses different independent variables / covariates / predictors / inputs, that will also be further explained here.

I have 33 colonies (Table 4.1) as presence records. The Department Environmental Affairs (DEA) gave the South African data which was verified through Google Maps, in longitude and latitude, converted into the correct format of decimal degrees. Lambert's Bay, Geyser Island and De Hoop areas are currently extinct areas, with zero as the colony number. The Namibian breeding locations' longitude and latitude coordinates were found through Google Mapping. Four of these locations: Mercury, Ichaboe, Halifax and Possession Islands have global Important Bird Area (IBA) status (Simmons et al. 1998). Counts are obtained from Earthwatch, source: Department Environmental Affairs (DEA) data. African penguin colony size is, however, not a true reflector of the habitat suitability. As previously discussed, penguins are social animals, grouping together, sometimes settling in worse habitat conditions than their surrounding areas. They are often sedentary, show a marked site fidelity, and do not leave the colonies for a long time. From the study, it is clear that locally dense records are in fact a true reflection of the relative suitability of the habitat, and therefore, the colonies are taken as the dependent variables.

Table 4.1: Penguin Colony Locations

Colony Location	x	y
Lambert's Bay	18.3267	-32.0978
Malgas Island	17.9254	-33.0528
Marcus Island	17.969535	-33.044076
Jutten Island	17.9554	-33.083392
Vondeling Island	17.9839	-33.153073
Dassen Island	18.0865	-33.4236
Robben Island	18.3723	-33.7998
Boulders	18.4513	-34.1972
Seal Island False Bay	18.5827	-34.13611
Stony Point	18.89333	-34.37222
Dyer Island	19.4148	-34.6841
Geyser Island	19.41369	-34.6897
De Hoop	20.5455	-34.4222
Jahleel Island	25.70474	-33.8055
Brenton Island	25.76518	-33.81796
St Croix Island	25.76972	-33.79944
Seal Island Algoa Bay	26.276379	-33.83382
Stag Island	26.283333	-33.833333
Bird Island	18.3026	-32.0901
Walvis Bay	14.50528	-22.9575
Hollam's Bird Island	14.516667	-24.633333
Sylvia Hill	14.8667	-25.15
Oyster Cliffs	14.8	-25.3333
Mercury Island	14.83283	-25.71942
Ichaboe Island	14.93333333	-26.28333333
Luderitz	15.13333	-26.65
Penguin Island	15.15452	-26.61623
Halifax Island	15.07977	-26.65099
North Reef	15.18389	-26.99639
Possession Island	15.19263	-27.01257
Pomona Island	15.25799	-27.19426
Plumpudding Island	15.53333	-27.63333
Sinclair Island	15.52033	-27.66529

After this, we get the African coastline, by reading the shapefile data using the shapefile function in the raster package, for the correct area under investigation:

```
1 coast = shapefile('GIS/GSHHS_f_L1_Africa.shp')
```

```
2 coastCoords = coast@polygons[[1]]@Polygons[[1]]@coords
3 coastCoords = data.frame(coastCoords)
4 names(coastCoords) = c('x','y')
5 coastCoords = coastCoords[coastCoords$x>11 & coastCoords$x<40
6 & coastCoords$y>c(-36) & coastCoords$y<c(-17),]
```

From line 1 in the above code, I get the shapefile of the African coast. From line 2, I extract the coordinates from "SpatialPolygonsDataFrame". I add this to a data frame in line 3, in the form of "x" and "y" columns. In line 5, I set the boundaries for the longitude or "x" coordinates, fixing the extent from 11 to 40 for "x" and from -36 to -17 for the latitude / "y" coordinates.

4.3.1 Pseudo-Absence / Background Data

Due to the lack of absence data in most data collections, some SDMs, including MaxEnt, make use of "pseudo-absence" records (background samples) to develop the model. Background data (e.g. Phillips et al. 2009) are attempting to characterize environments in the study region, as well as establish the environmental domain of the study. To describe the concept of "pseudo-absences", one should know it is also used for generating the non-presence class for logistic models. However, in this case, we must try to estimate where absences might occur. The whole area may be sampled, except at presence locations, or places unlikely to be suitable for the species may be sampled. The background concept is preferred, because it requires fewer assumptions. It also has some coherent statistical methods for dealing with the "overlap" between presence and background points (e.g. Ward et al. 2009; Phillips and Elith, 2011).

Elith et al. (2006) completed an in depth study to show how different SDM methods predicted species' distributions. Presence-only data with pseudo-absences (a random sample of 10 000 sites from each region) were used for training the models and presence-absence data were used for evaluations. They found that the presence-only data were effective in modelling the species' distributions. It is possible that pseudo-absences will coincide with presence records, especially when randomly choosing these pseudo absences from the study region. This is, however, widely accepted across different models and did not negatively impact the outcome. Background points that are too close to presences can give false projections. When one has few presences, it is better to select points at least 2 degrees away from any presence point (Barbet-Massin et al. 2012). Ideally data collections should strive to collect both presence and absence data, making the modelling more robust and accurate. Unfortunately this is not always possible, and absence data could even be misleading due to species and their environment not being in equilibrium, or the species might be difficult to detect. A few SDM methods (GARP, MaxEnt and Ecological Niche Factor Analysis (ENFA)) have been developed to deal with data sets lacking accurate absence data.

MaxEnt's default setting sampled 10 000 background points from the study area (linear, coastline.) The code will be shown in the next section.

4.4 Environmental Data

Great consideration should be taken when choosing the predictor variables for a model (discussed in Chapter 3), as they are a primary decision maker of how the model output is formulated. It is important to choose variables that are relatively independent from each other, relevant to the dynamics of the study species and to the resolution of the study (Pearson et al., 2004). To build a meaningful model for a species' distribution, you require knowledge regarding the species' biology, population dynamics, sensitivity to human disturbances, etc. Choosing more predictor variables would not increase the chances of a successful outcome (5-10 variables are considered ideal) and the balance of predictor variables should depend on the spatial scale being considered. Variables that have a direct impact on a species' distribution should be used above indirect variables, as the former is more relevant.

It is essential to rely on a priori knowledge of which variables to include or exclude (Elith and Leathwick, 2009; Huntley et al., 2008). Refer to Chapter 3: Demography of the African Penguin, specifically Section 3.6, to understand which variables I chose and why.

It is possible to start a SDM process, with all the variables available, and rely on the SDM's outcome to tell you the variables' contributions and accordingly eliminate variables. This procedure, however, can not replace one where a prior selection is built on existing knowledge and theory. On the other hand, if only a priori knowledge is used, a relationship is forced between the species' probability of occurrence and a climatic variable. When allowing the model to eliminate variables, additional relationships with previously unexpected variables, which may be important for the species' distribution, can be discovered, possibly leading to new knowledge regarding the species' habitat.

We also implement the environmental layers (rasters). This is needed to obtain the suitability of the areas investigated. In order to do this, this code is used:

```
envFiles <- list.files("GIS/")
envFiles <- envFiles[c(grep('jpg', envFiles), grep('tif', envFiles))]
envFiles = envFiles[-grep('xml',envFiles)
envNames = {}
for(r in 1:length(envFiles)){
  rast <- raster(paste("GIS/",envFiles[
  rast = crop(rast, extent(c(11,40,-36,-17)))
  assign(names(rast), rast)
  envNames[r] = names(rast)
  if(r==1){
    envStack = stack(rast)
  } else if(r>1 & r<=10){
    envStack = stack(envStack, rast)
  }
```

```
6  rm(rast)
}
```

From the above code, line 1 names of all the files in the GIS directory. The description below explains which layers I used and where I extracted them from. From line 2, I only select the jpg and tif files. I also deselect the xml files (line 3). Line 4 creates a vector to store the names of the rasters. In line 5, I loop through the raster file names, open, crop, assign and place them in the raster stack. In line 6, I remove the placeholder raster.

I now plot the raster layers by using the code below. The grids, in the form of rasters, do not completely overlap. They must all have the same geographic bounds and cell size (i.e. all the file headings must match each other perfectly). The function "fix-extent" is used to make them align and add them to the raster stack.

```
#plot(envStack)

envStack2 = fix_extent(bio1, envStack)
envStack = addLayer(envStack2[[1]], bio1)
envStack2 = fix_extent(bio5, envStack)
envStack = addLayer(envStack2[[1]], bio5)
envStack2 = fix_extent(bio12, envStack)
envStack = addLayer(envStack2[[1]], bio12)
envStack2 = fix_extent(bio15, envStack)
envStack = addLayer(envStack2[[1]], bio15)
#envStack2 = fix_extent(bio11, envStack) #additional layers for seasonality
#envStack = addLayer(envStack2[[1]], bio11)
#envStack2 = fix_extent(bio13, envStack)
#envStack = addLayer(envStack2[[1]], bio13)
#envStack2 = fix_extent(bio14, envStack)
#envStack = addLayer(envStack2[[1]], bio14)
```

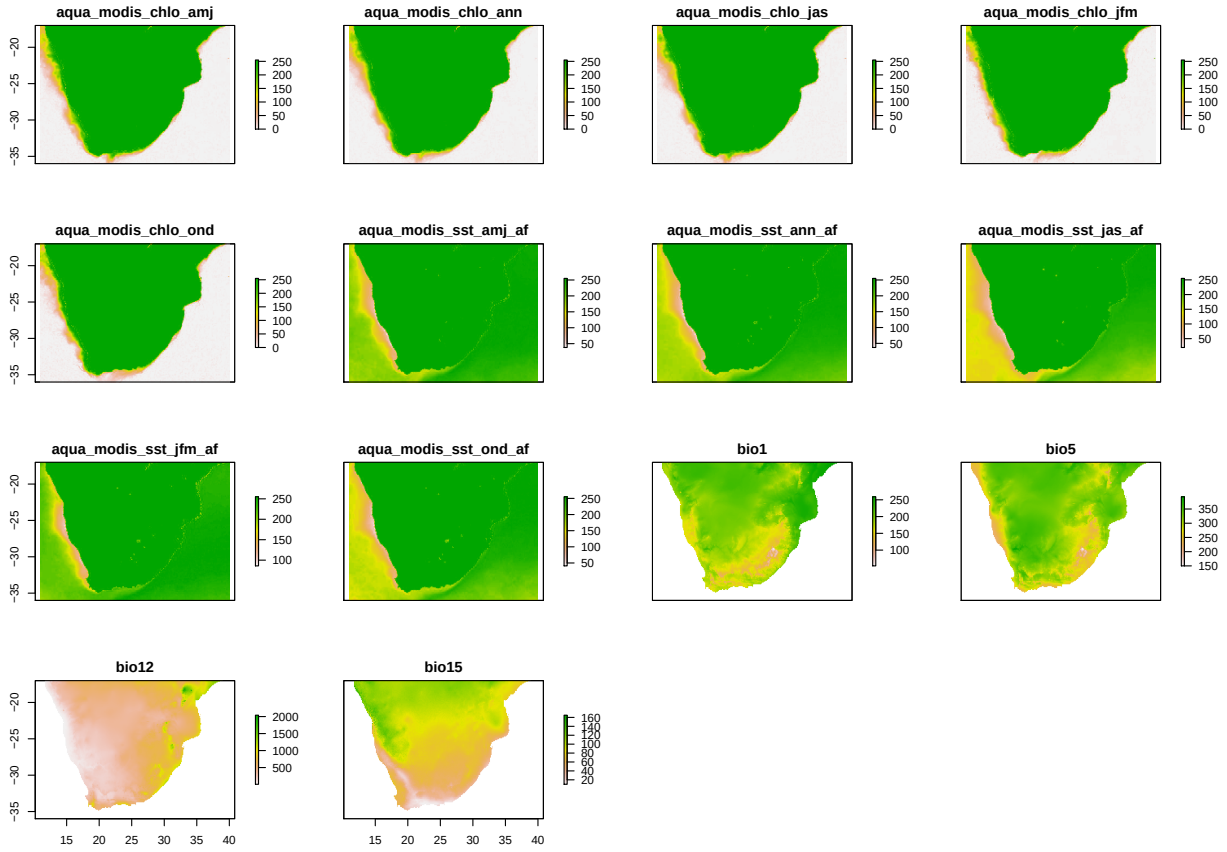

Figure 4.3: Environmental data in raster format, where 'chlo' stands for chlorophyll ($\frac{mg}{m^3}$) and 'sst' for sea surface temperature (degrees celcius), 'amj' stands for spring, 'ann' for annual, 'jas' for summer, 'jfm' for winter and 'ond' for autumn from African Marine Atlas. Refer to Table 4.2 for the bioclimatic variable description and units.

I now move the points on land to the nearest cell in the sea and the points in the sea to the nearest cell in land, so that everything is included. I also add the fields required by biogeo. All environmental rasters now have the same geographic bounds, cell size and same column file headings. "Checkdatastr" checks the data structure to see which of the required columns are missing from the dataframe. Now we have the presence data.

```
checkdatastr(pres)
pres$Species = 'Penguin'
pres = addmainfields(pres, species = 'Species')
#presLand = nearestcell(pres, bio1)
presLand = pres
```

We now do the same for all coastal coordinates and use these as background points:

```
back = coastCoords
checkdatastr(back)
```

```
back$Species = 'Penguins'
back = addmainfields(back, species = 'Species')
```

To extract (from the "raster" package) the environmental data, the following code is run:

```
presEnv = data.frame(extract(envStack, cbind(pres$x, pres$y)))
presEnvLand = data.frame(extract(envStack, cbind(pres$x_land, pres$y_land)))
#write.csv(presEnvLand, file = "presEnvLand.csv")
```

I replaced any NA values from points in the sea, with values using land coordinates, by using the below code (for presence and background data). The generic function "is.na" shows which elements are missing. "arr.ind" indicates which indices are true, when called up by adding "=T" (true).

```
presEnv[is.na(presEnv)] = presEnvLand[which(is.na(presEnv), arr.ind=T)]
backEnv = data.frame(extract(envStack, cbind(back$x_land, back$y_land)))
backEnvLand = data.frame(extract(envStack, cbind(back$x_land, back$y_land)))
#write.csv(backEnvLand, file = "backEnvLand.csv")
backEnv[is.na(backEnv)] = backEnvLand[which(is.na(backEnv), arr.ind=T)]
```

I now add the environmental data to the coordinate data:

```
pres = cbind(pres[,c('location','x','y','x_land','y_land')], presEnv)
back = cbind(back[,c('x_land','y_land','x_land','y_land')], backEnv)
#write.csv(pres, file = "pres_presEnv.csv") #to write in .csv format
#write.csv(back, file = "back_backEnvLand.csv")
```

The raster for presence and background data is written in the correct manner.

```
rasID = envStack[[1]]
rasID[] = 1:length(rasID)
writeRaster(rasID, 'rasID.tif', overwrite=T)
pres$ID = extract(rasID, pres[,c('x','y')])
pres$IDland = extract(rasID, pres[,c('x_land','y_land')])
back$ID = extract(rasID, back[,c('x_land','y_land')])
back$IDland = extract(rasID, back[,c('x_land','y_land')]
```

Any duplicated points are removed by this code:

```
pres = pres[!duplicated(pres$ID) & !duplicated(pres$IDland),]
back = back[!duplicated(back$ID) & !duplicated(back$IDland),]
```

Using this code, I remove the "presence coordinates" from the "background coordinates":

```
back = back[!back$ID%in%c(pres$ID,pres$IDland) & !back$IDland%in%c(pres$ID,pres$IDland),]
```

To check for the collinearity among environmental predictors, this code is used:

```
envNames = names(envStack)
names(back)
colnames(back)[3] = "x"
colnames(back)[4]= "y"
names(back)

dat = rbind(pres[,-1],back[,names(back)%in%names(pres)])
M <- cor(dat[,envNames])

corrplot.mixed(M, lower = "circle", upper = "number", tl.pos = c("lt"))
```

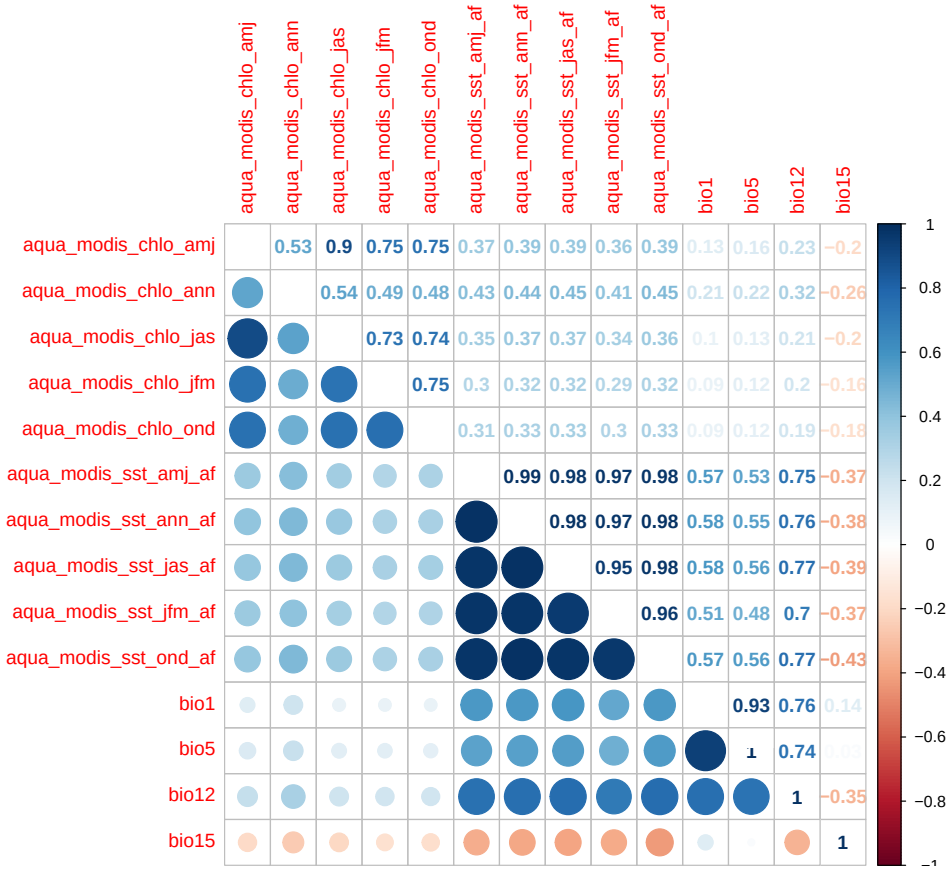

Figure 4.4: Checking collinearity among environmental predictors. The red dots indicate negative correlations, whilst the blue dots show positive correlations. The strength of the correlation is indicated by dot size. 'chlo' stands for chlorophyll ($\frac{mg}{m^3}$) and 'sst' for sea surface temperature (degrees celcius), 'amj' stands for spring, 'ann' for annual, 'jas' for summer, 'jfm' for winter and 'ond' for autumn from African Marine Atlas. Refer to Table 4.2 for the bioclimatic variable description and units.

I will now indicate the source and description of the explanatory variables (also called independent variables), which explain changes in the response variable (dependent variable). I extracted the variables from Bioclim (see Table 4.2) : annual mean air temperature (for seasonality: mean temperature of warmest (summer); coldest (winter) quarter), annual precipitation (for seasonality: precipitation of driest (summer); wettest (winter) month), and precipitation seasonality (coefficient of variation). Bioclim (www.worldclim.org/bioclim) data is obtained from WorldClim (www.wordlclim.org, Hijmans et al., 2004). Table 4.2 shows the bioclimatic variables, with scaling factors and units. It is a set of global climate layers (gridded climate data) with a high spatial resolution. The data is used for mapping and spatial modelling.

Table 4.2: Bioclimatic Variables

Label	Variable	Scaling Factor	Units
BIO1	Annual Mean Temperature	10	Degrees Celsius
BIO2	Mean Diurnal Range (Mean of monthly)	10	Degrees Celsius
BIO3	Isothermality (BIO2/BIO7)	100	Degrees Celsius
BIO4	Temperature Seasonality (Standard Deviation)	100	Degrees Celsius
BIO5	Max Temperature of Warmest Month	10	Degrees Celsius
BIO6	Min Temperature of Coldest Month	10	Degrees Celsius
BIO7	Temperature Annual Range (BIO5-BIO6)	10	Degrees Celsius
BIO8	Mean Temperature of Wettest Quarter	10	Degrees Celsius
BIO9	Mean Temperature of Driest Quarter	10	Degrees Celsius
BIO10	Mean Temperature of Warmest Quarter	10	Degrees Celsius
BIO11	Mean Temperature of Coldest Quarter	10	Degrees Celsius
BIO12	Annual Precipitation	1	Millimetres
BIO13	Precipitation of Wettest Month	1	Millimetres
BIO14	Precipitation of Driest Month	1	Millimetres
BIO15	Precipitation Seasonality (Coefficient of Variation)	100	Fraction
BIO16	Precipitation of Wettest Quarter	1	Millimetres
BIO17	Precipitation of Driest Quarter	1	Millimetres
BIO18	Precipitation of Warmest Quarter	1	Millimetres
BIO19	Precipitation of Coldest Quarter	1	Millimetres

One can download the variables for different spatial resolutions, from 30 seconds (about 1 km^2) to 10 minutes (about 340 km^2). Each download is a 'zip' file containing 12 GeoTiff (.tif) files, one for each month of the year (January is 1; December is 12). I used 2.5-minutes (of a longitude/latitude degree) spatial resolution (this is about 4.5 km at the equator). This is suitable for modelling purposes, as it is close to the African penguin's mean foraging distance.

WorldClim version 2 has average monthly climate data for minimum, mean, and maximum temperature, and for precipitation for years 1970 to 2000. (WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, Fick, S.E. and R.J. Hijmans, 2017.)

The variables discussed are derived from the monthly temperature and rainfall values, in order to generate meaningful variables. The bioclimatic variables represent annual data (e.g. annual precipitation), seasonality (e.g. annual range in precipitation) and limiting or extreme environmental factors (e.g. precipitation of the wet and dry quarters, where a quarter is a period of three months or 1/4 of the year).

I also extracted sea surface temperature data and chlorophyll counts (annual, and for seasonality: summer and winter data) from the African Marine Atlas (http://omap.africanmarineatlas.org/index.htm).

When I download it, four files are generated and I use the .jpg files. MODIS-Aqua Chlorophyll data is described as a 4km-resolution data set that consists of satellite measurements of global ocean colour and sea surface temperature (SST) data, obtained by the Moderate Resolution Imaging Spectroradiometer (MODIS), in orbit on the Aqua (formerly EOS PM) platform. MODIS ocean colour and SST products are processed and distributed by the Ocean Biology Processing Group (OBPG). (Source: US NASA Ocean Color Website, Feldman, G. C., C. R. McClain, Ocean Color Web, MODIS-Aqua, NASA Goddard Space Flight Center. Eds. Kuring, N., Bailey, S. W. June 2006, http://oceancolor.gsfc.nasa.gov/).

MODIS-Aqua SST, sea surface temperature, is a data set that consists of satellite measurements of the sea surface temperature (SST) obtained by the Moderate Resolution Imaging Spectroradiometer (MODIS) in orbit on the Aqua platform. This SST product is processed and distributed by the Ocean Biology Processing Group (OBPG). The generation of SST product from the MODIS sensors is performed using the Multi-Sensor Level-1 to Level-2 software (MS112), based on algorithms and logic originally developed by the Rosenstiel School of Marine and Atmospheric Science (RSMAS) at the Universty of Miami. Source: Ocean Colour Web.

The contribution of each explanatory variable will be shown in the Hierarchical Partitioning outputs, including seasonality, in Chapter 5: Results.

4.5 Fish Stock Assessment

The Department of Agriculture, Forestry and Fisheries (DAFF) conducted their 33^{rd} consecutive, annual, November biomass survey over a 52 day period, on board of the FRS Africana in 2016. From their information, they found, that 71% of the biomass of sardine, 183.4 thousand tonnes, occur in the area to the west of Cape Agulhas (see Figure 4.5). This has increased since 2015, mainly due to a large influx of younger fish to the west coast in 2016. 77% of the proportion of the biomass of anchovy occur in the area to the west of Cape Agulhas (see Figure 4.6). This is the highest it has been since 1995, as well as before the shift of anchovy to the South Coast in 1996. Figure 4.7 indicates the relative percentages of sardine and anchovy found to the west and east of Cape Agulhas. Due to this data, the suggested eastward shift for this species, and mechanisms for maintaining such should be re-investigated (DAFF, 2016). The low biomass of sardine is a concern, especially in the area discussed, if future occurrence of sardine is predominantly dependent on successful west coast spawning.

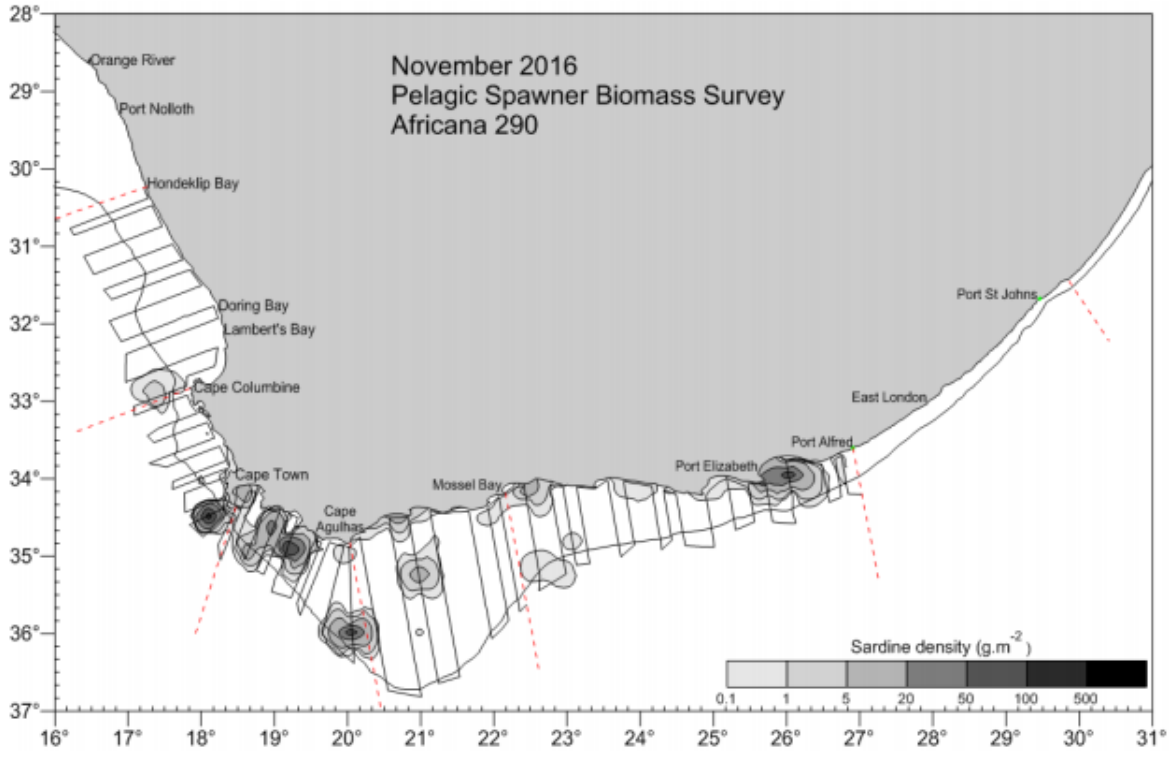

Figure 4.5: Distribution and relative density of sardine (Department of Agriculture, Forestry and Fisheries (DAFF), 2016).

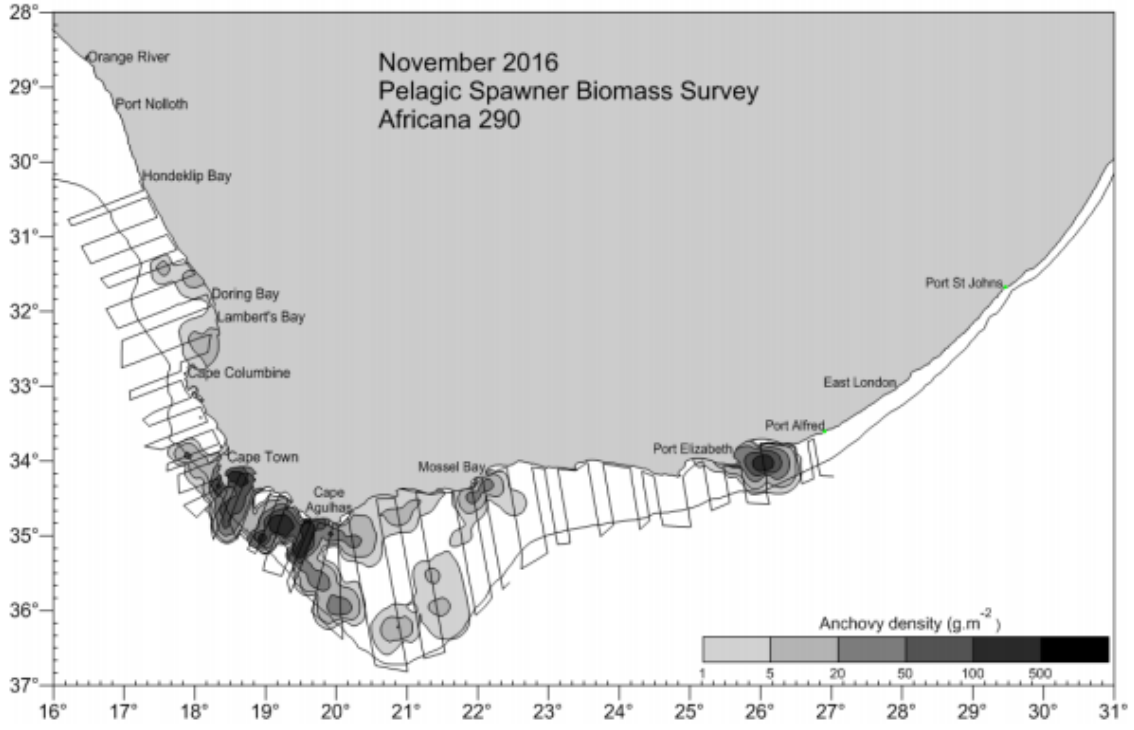

Figure 4.6: Distribution and relative density of anchovy (DAFF, 2016).

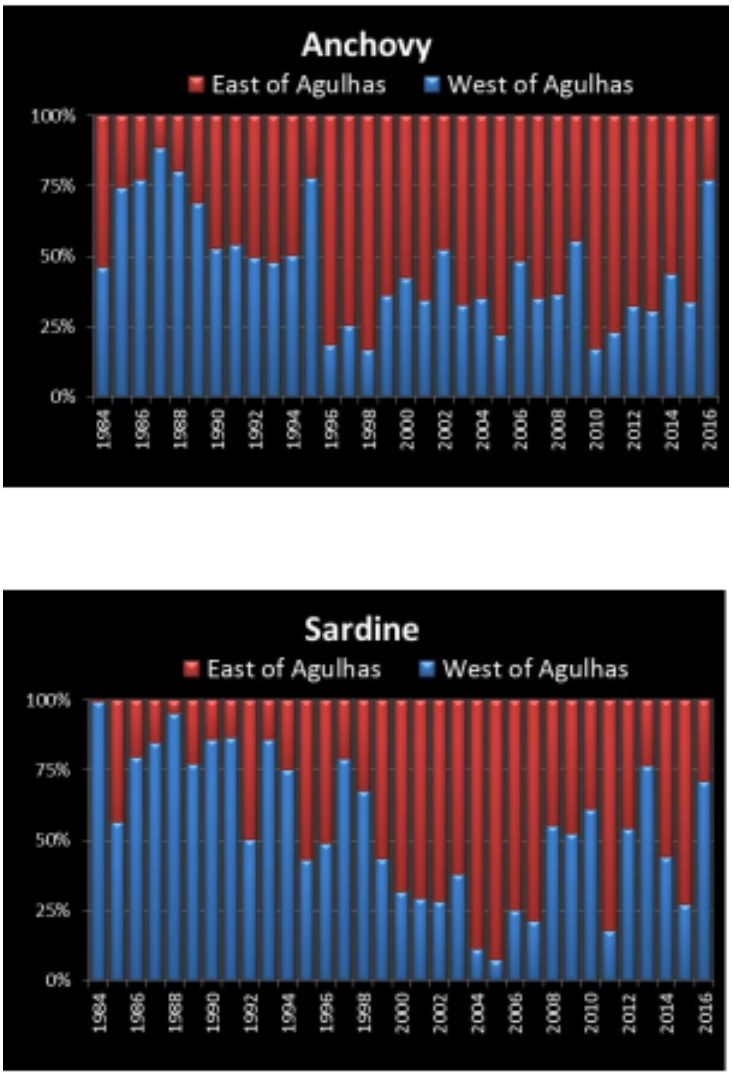

Figure 4.7: Relative percentage of the biomass found to the west and east of Cape Agulhas, with anchovy indicated above and sardine below (DAFF, 2016).

4.6 Modelling Methods and Validation

It is crucial to evaluate the model with independent data, to ensure that overfitting does not occur. Overfitting would lead to deceptive conclusions on the role of the predictor variables, as well as their relationships with the study species. There are various ways of obtaining such an independent data set. The original data can be split into a separate training and test set using random split, spatial split or cross validation resampling methods. If independent data is available, it can be used for testing. This data could be collected independently from the first data set, or it can be temporally or spatially independent data. Bahn and McGill (2013) states that a truly independent and spatially segregated

data set is necessary, in order to test if the model can be used to make predictions in new areas / environments.

For evaluating using cross-validation, the percentage training data used is 70% and for the testing model, it is 30%, as seen in the code below.

```
p = pres[,c('x','y')]
a = back[,c('x','y')]
set.seed(10) # random number generator where the values are unique to that seed
pSamp = sample(x = c(1:nrow(p)), size=round(nrow(p)*0.7,0), replace=F)
ptrain = p[pSamp,]
ptest = p[-pSamp,]
aSamp = sample(x = c(1:nrow(a)), size=round(nrow(a)*0.7,0), replace=F)
atrain = a[aSamp,]
atest = a[-aSamp,]

#Full model with all occurrences:
1 betaRCs <- seq(0.05,0.95,0.05)
2 bestAUC = 10
3 names(envStack)
4 keeps = c('aqua_modis_chlo_ann','aqua_modis_sst_ann_af','bio1', 'bio12', 'bio15')
#Select which variables to keep
5 mxMod = maxent(x = dat[,keeps],
                 p = c(rep(1,nrow(pres)), rep(0,nrow(back))),
                 args = c('responsecurves',
                 '-p','nothreshold',paste0('beta_lqp=',betaRCs[bestAUC]),
                 paste0('beta_hinge=',betaRCs[bestAUC])),
                 removeDuplicates=T)
```

As seen in line 4 in the code, we choose which environmental variables to keep. Response curve outputs are added ("responsecurves" in line 5) to the html output file containing all the outputs. The more complex product and threshold features were removed from MaxEnt, to avoid the danger of overfitting. What is left is linear, quadratic and hinge features. The regularization parameter is specified for the MaxEnt function, also in order to remove duplicate occurences within the grid cells.

4.6.1 Area Under Curve (AUC)

The area under the receiver operating curve (AUC) was used to evaluate the discriminative ability of MaxEnt models. Area Under Curve (AUC) remains one of the most widely used and unbiased measures of accuracy (Pearson et al., 2004). The AUC is obtained from the receiver operating curve (ROC), which depicts the relationship between the proportion of true positives on the y-axis (sensitivity, profits) and false positives on the x-axis (1 - specificity, expenses), with varying probability

thresholds. The true positive rate is defined as $\frac{Positives \quad correctly \quad classified}{Total \quad positives}$ and false positive rate as $\frac{Negatives \quad incorrectly \quad classified}{Total \quad negatives}$ (eg. Rivera et. al. 2017). Sensitivity is the percentage of actual presences predicted, which is 1 - false negative or omission rate. It thus quantifies omission errors. Sensitivity (true positive rate) is plotted against 1 - specificity. Specificity is the percentage of actual absences predicted or true negative rate. See Table 4.3 for the contingency table.

Table 4.3: Contingency Table for a Given Threshold

		Actual Value (Data)	
		Presence (pos)	Absence (neg)
Predicted Outcome (Model)	Presence (pos)	True Positive (TP)	False Positive (FP)
	Absence (neg)	False Negative (FN)	True Negative (TN)

AUC thus measures the ability of predictions to discriminate between presences and absences (Elith and Graham, 2009). Good model performance is characterized by large areas under the ROC curves, hence a curve that maximizes sensitivity for low values of 1 - specificity. AUC ranges from 0 to 1, with 1 being a model with perfect discrimination between presences and absences, and 0 to 0.5 suggesting that the model is no better than a random model. The closer the curve gets to the upper left corner, maximizing sensitivity for low values of 1-specificity, the greater the AUC and the better the model's prediction of reality. It measures the likelihood that a randomly selected presence point is located in a raster cell with a higher probability value for species occurrence, than a randomly selected absence point. As shown in Figures 4.8 to 4.10, annual (AUC of 0.889) and the seasons (summer AUC 0.898; winter AUC 0.888) performed well. The red, training data line indicates the "fit" of the model to the training data.

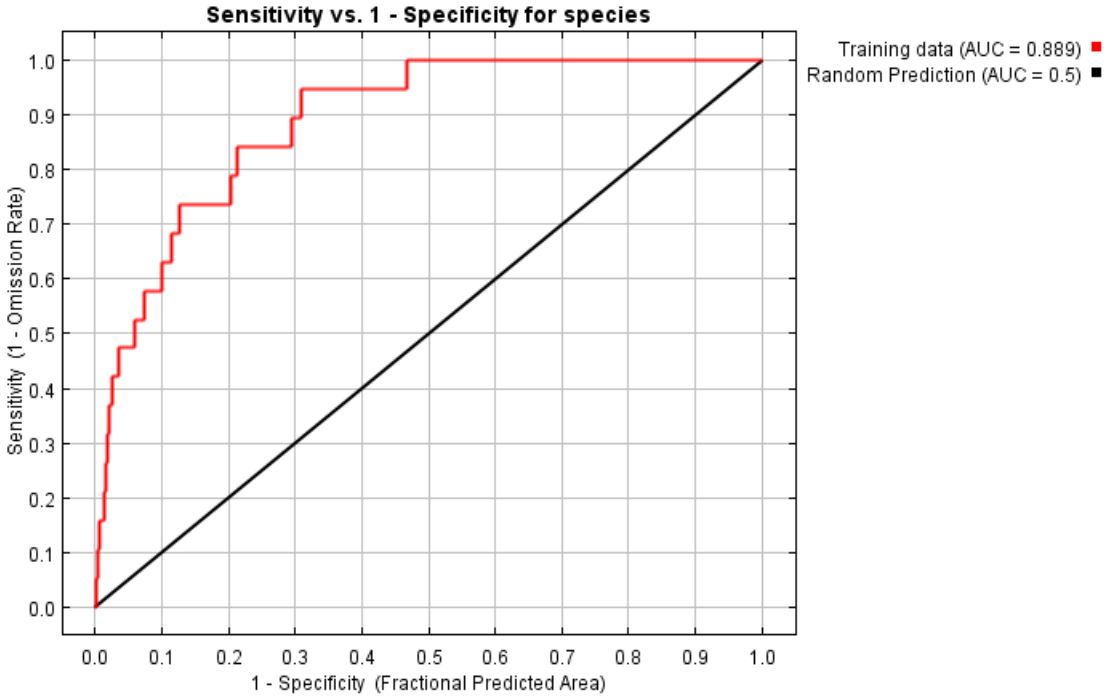

Figure 4.8: AUC annual data.

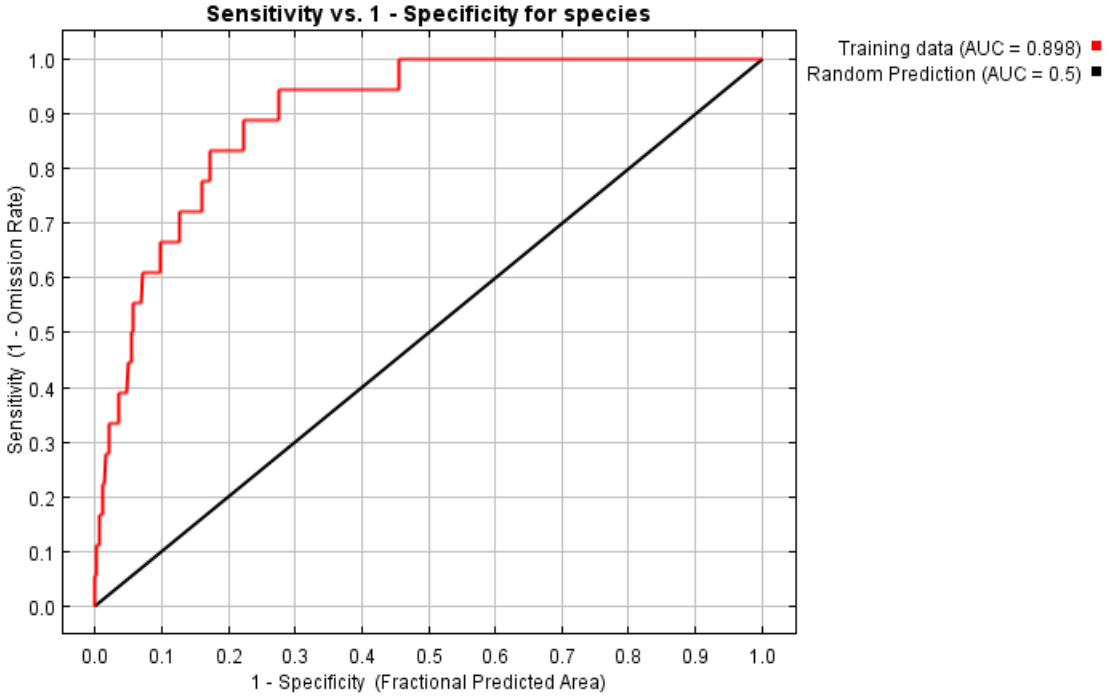

Figure 4.9: AUC summer data.

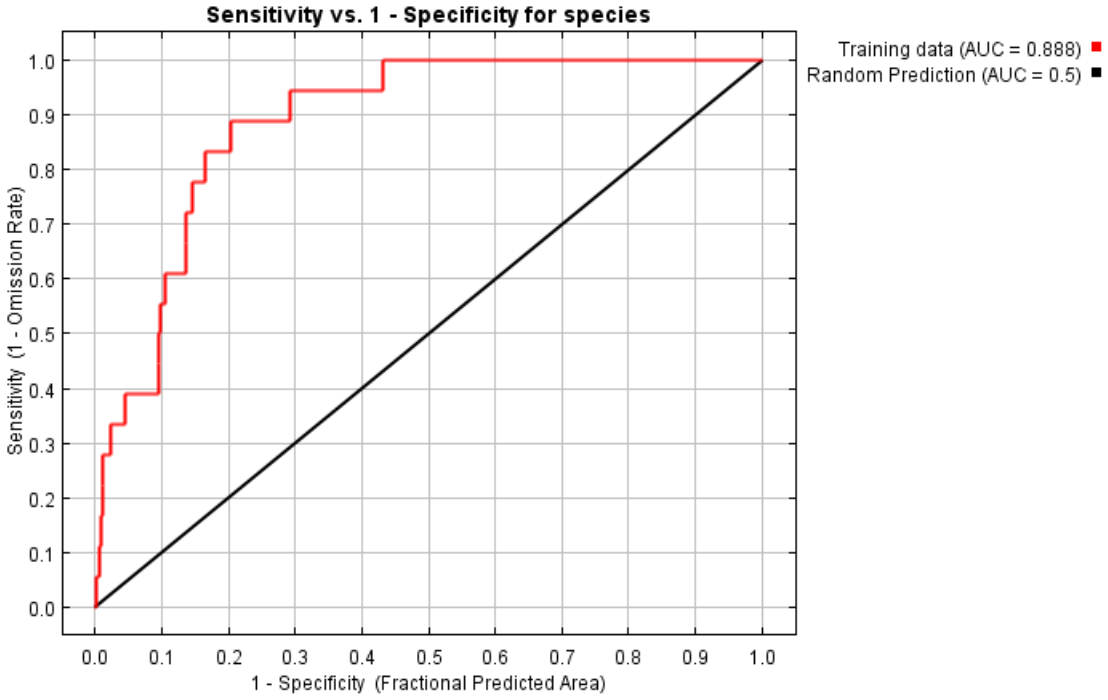

Figure 4.10: AUC winter data.

4.6.2 Interpretation of ROC and AUC for Model Evaluation

Using the ROC analysis, as used in my model, has the main advantage that the area under the ROC curve (AUC) provides a single measure of model performance, independent of any particular choice of threshold. The thresholds in my models produce good maps of the species' potential distributions according to the environmental variables chosen, the areas of suitable environmental conditions.

The ROC curve is obtained by plotting sensitivity on the y axis and 1-specificity on the x axis, for all possible thresholds, as previously explained in Chapter 4.6.1. For a continuous prediction, the ROC curve typically contains one point for each test instance. For a discrete prediction, in addition to the origin, there will typically be one point for each of the different predicted values. The area under the curve (AUC) is usually obtained by connecting the points with straight lines. This is called the trapezoid method. Parametric methods would have fit a curve to the points. The AUC has an instinctive interpretation, that is the probability that a random positive instance and a random negative instance are correctly interpreted. This interpretation shows that the AUC is not sensitive to the relative numbers of positive and negative instances in the test data set.

When only presence data are available, as in my instance, it would come across that ROC curves are inapplicable, because without absences, there seems to be no source of negative instances with which to measure specificity. However, we can avoid this problem by regarding a different classification problem. This is the task of distinguishing presence from random, rather than presence from absence. We state that for each pixel x in the study area, we define a negative instance x_{random}. We also define a positive instance $x_{presence}$ for each pixel x that is included in the species' true ge-

ographic distribution. Without seeing the labels random or presence, a species distribution model can then make predictions for the pixels corresponding to these cases. We can thus make predictions for both a sample of positive instances (the presence localities) and a sample of negative instances (background pixels chosen uniformly at random, or according to another background distribution). Together these are enough to define a ROC curve, which can then be analysed with all the standard statistical methods of ROC analysis. This process can be interpreted as using pseudo-absence instead of absence in the ROC analysis, such as used in Wiley et al. (2003). With presence-only data, as in my model's case, the maximum achievable AUC is less than 1 (Wiley et al., 2003).

To clarify why my model performs well with good AUC values and gives a good, continuous prediction, it is because of what it is based on. The environmental variables chosen in my model are: sea surface temperature, mean temperature, chlorophyll count, precipitation and precipitation seasonality (coefficient of variation).

Chapter 5

Results

5.1 Hierarchical Partitioning

To assess which environmental variables are making the greatest contribution to the model, these variables that are contributing to fitting the model are being tracked, while the MaxEnt model is being trained. Each step of the MaxEnt algorithm increases the gain of the model. This is performed by modifying the coefficient for a single feature. The increase in the gain is assigned to the environmental variables that the feature depends on. The percent contribution is obtained by converting to percentages at the end of the training process, and is shown as the Hierarchical Partitioning figures.

Hierarchical Partitioning (HP) is a method to assess the contribution of each environmental variable. It is an analytical method of multiple regression that identifies the most likely causal factors, while alleviating multicollinearity problems (Olea et al. 2010). Its use is increasing in ecology and conservation, as the ranking obtained in HP is being used as a criterion for establishing priorities of conservation. HP is important for predicting the response of biodiversity to climate change. The maximum likelihood method is used to fit the model to data in HP (i.e. argument of goodness-of-fit "logLik"). Likelihood methods are used to gain information from incomplete observations (for example, Fienberg et.al. 2012; Dempster et. al. 1977).

In Figures 5.1, 5.2 and 5.3 (annual, summer and winter), the relative contributions of the 5 environmental variables are shown. Please refer to Chapter 3.6 for the explanation on chosen variables. Sea surface temperature is the biggest contributing factor (72.4%, 53.2% and 46.9% respectively), followed by mean land / air temperature (14.5%, 37.3% and 40%.)

As an explanation, in each iteration of the training algorithm, the increase in regularized gain is added to the contribution of the corresponding variable to determine the first estimate. The increase in regularized gain, is however subtracted, if the change to the absolute value of the regularization parameter, lambda, is negative. Each variable's value on training presence and background data is randomly permuted for the second estimate. After the model is re-evaluated on the permuted data, the resulting drop in training AUC, normalized to percentages, is indicated (see Tables 5.1 to 5.3). The algorithm converged after 360 iterations.

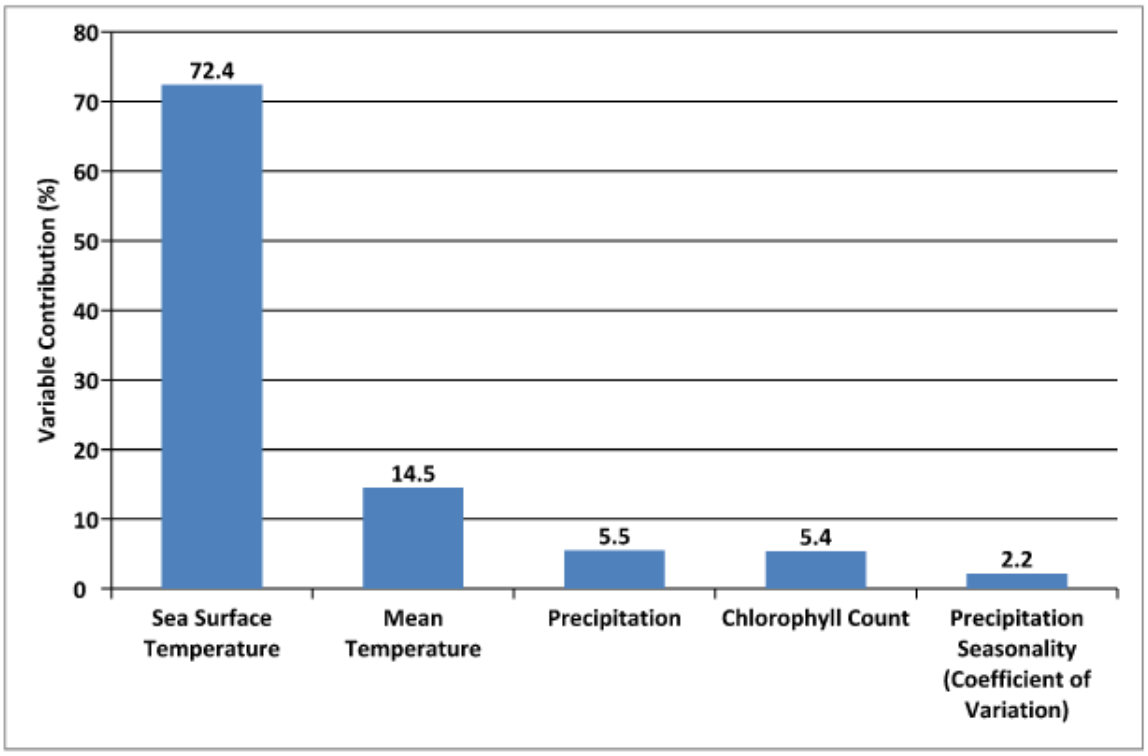

Figure 5.1: Annual Hierarchical Partitioning values.

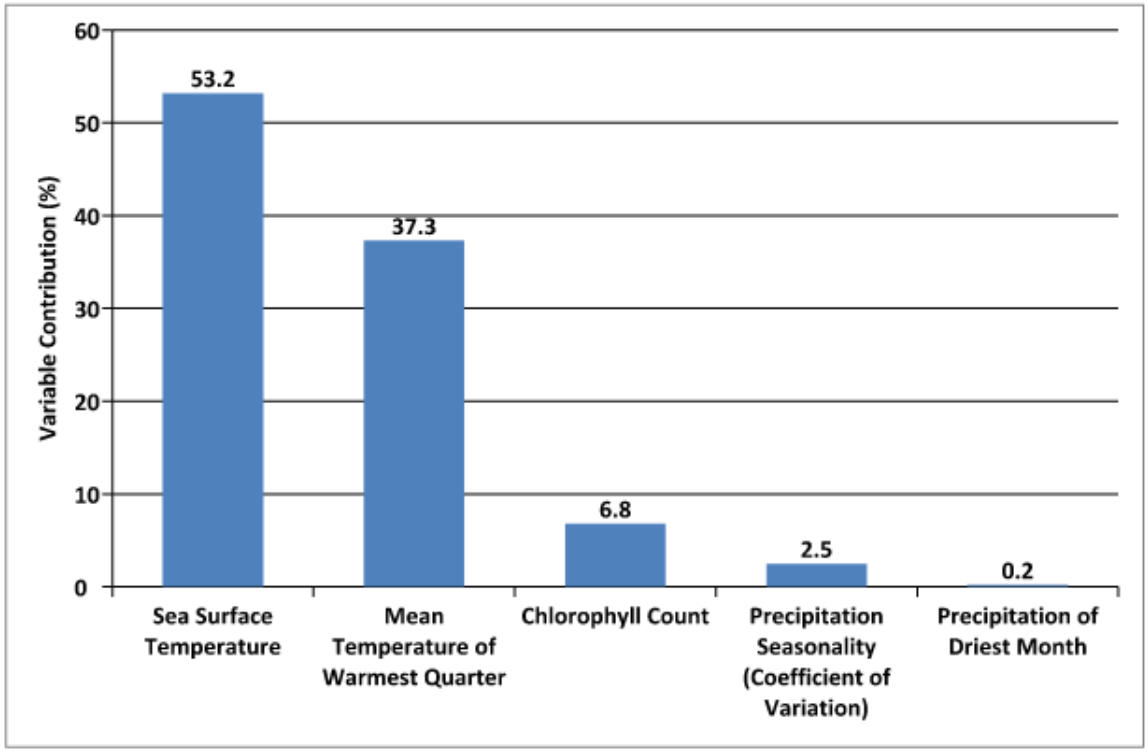

Figure 5.2: Summer Hierarchical Partitioning values.

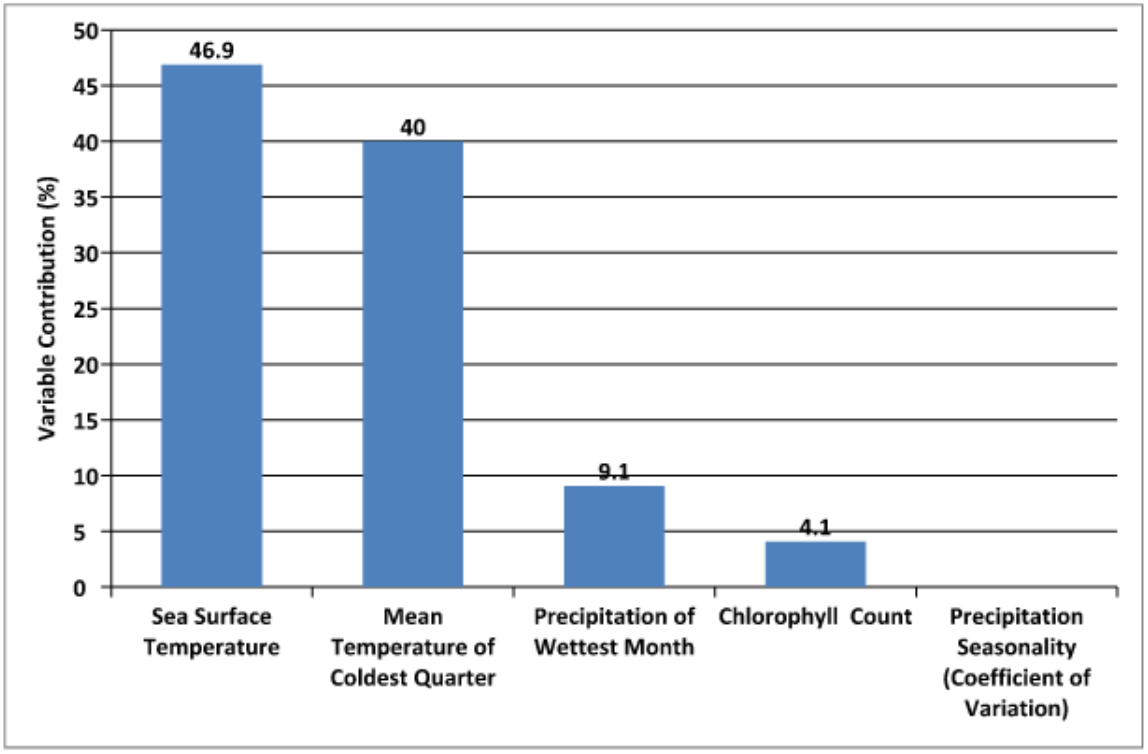

Figure 5.3: Winter Hierarchical Partitioning values.

Table 5.1: Annual Variable Importance: where annual SST is the highest percent contributor, as well as the highest permutation importance.

Variable	Percent Contribution	Permutation Importance
Annual SST	72.4	58.7
bio1 (Annual Mean Temperature)	14.5	9.7
bio12 (Annual Percipitation)	5.5	20.1
Annual Chlorophyll	5.4	6
bio15 (Percipitation Seasonality)	2.2	5.5

Table 5.2: Summer Variable Importance: where summer SST is the highest percent contributor, however bio10 (Mean Temperature of Warmest Quarter) shows the highest permutation importance.

Variable	Percent Contribution	Permutation Importance
Summer SST	53.2	14.2
bio10 (Mean Temperature of Warmest Quarter)	37.3	73.1
Summer Chlorophyll	6.8	5.6
bio15 (Percipitation Seasonality)	2.5	4.4
bio14 (Percipitation of Driest Month)	0.2	2.7

Table 5.3: Winter Variable Importance: where winter SST is the highest percent contributor, however bio11 (Mean Temperature of Coldest Quarter) shows the highest permutation importance.

Variable	Percent Contribution	Permutation Importance
Winter SST	46.9	10.3
bio11 (Mean Temperature of Coldest Quarter)	40	75.4
bio13 (Percipitation of Warmest Month)	9.1	11.6
Winter Chlorophyll	4.1	2.7
bio15 (Percipitation Seasonality)	0	0

5.1.1 Jackknife Test of Variable Importance

The jackknife test is performed to get an alternate estimate of which variables are most important in the model. For this test, each variable is excluded in turn, and a model is created with the remaining variables. Another model is obtained using each variable in isolation. Then a model is created using all the variables.

I will now discuss the effect of the regularization parameter. Regularization is a process of introducing additional information to the objective function, in order to prevent overfitting. It can be seen as a smoothing factor. It controls the model complexity, the right model fit in between the simple and complex models, so that the model is better at predicting. A too simple model will be a very poor generalization of data, whilst a too complex model may not perform well with test data, due to overfitting. The regularization parameter, known as lambda, which is an input to the model, reduces overfitting. It reduces the variance of the estimated regression parameters. However, it does this at the expense of adding bias to the estimate. Increasing lambda results in less overfitting, but also greater bias, thus careful assessment needs to be performed.

The environmental variable with highest gain, when used in isolation, is aqua-modis-sst-ann-af (annual SST, the longest dark blue bar, in Figure 5.4), which therefore appears to have the most useful information by itself. It allows a good fit to the training data. We also see that if MaxEnt uses only

bio15 (precipitation seasonality, the shortest dark blue bar in Figure 5.4) it achieves almost no gain, so that variable is not, by itself, useful for estimating the habitat suitability of African penguins. As seen from shortest lighter blue bars in Figure 5.4, the environmental variable that decreases the gain the most when it is omitted, is annual SST, which therefore appears to have the most information that is not present in the other variables. The light blue bars in Figure 5.4 are never longer than the red bar, which shows that predictive performance does not improve when the corresponding variables are not used. Therefore, the 5 environmental variables show a good choice of variable selection.

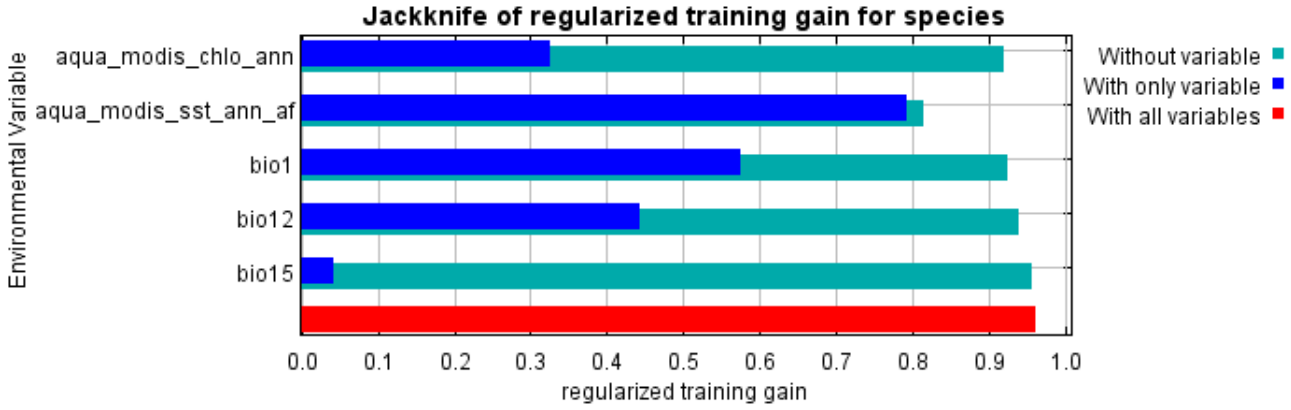

Figure 5.4: Jackknife test from MaxEnt on the annual dataset.

Figure 5.5 shows the results of the jackknife test of variable importance for summer. The environmental variable with the highest gain when used in isolation is bio10 (mean temperature of warmest quarter), which therefore appears to have the most useful information by itself. Bio10, is also the environmental variable that decreases the gain the most when it is omitted (as seen from shortest lighter blue bar), which therefore appears to have the most information that is not present in the other variables.

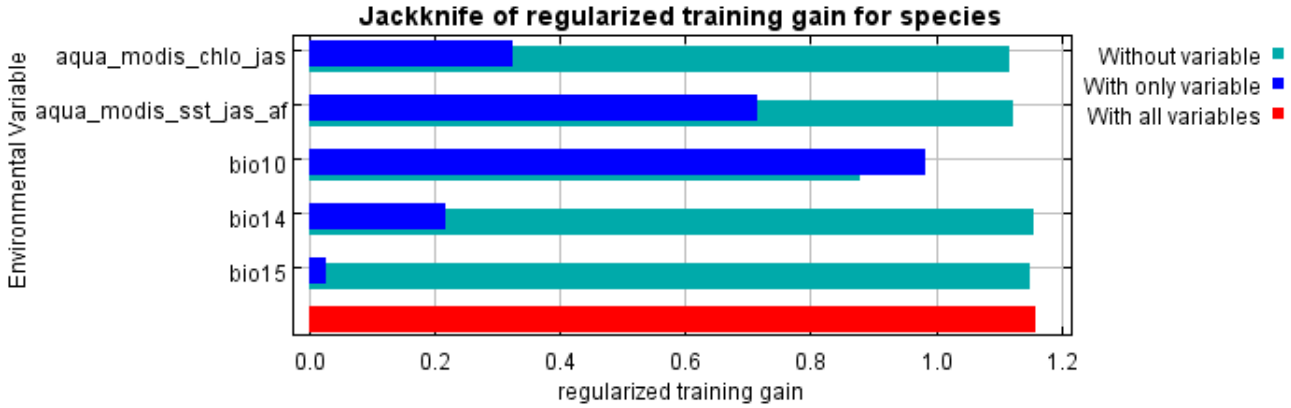

Figure 5.5: Jackknife test from MaxEnt on the summer dataset.

For the winter jackknife test of variable importance (see Figure 5.6, longest dark blue bar), the environmental variable with highest gain when used in isolation is aqua-modis-sst-jfm-af (winter sea surface temperature), which therefore appears to have the most useful information by itself. The environmental variable that decreases the gain the most when it is omitted is bio11 (mean temperature of coldest quarter, shortest light blue bar), which therefore appears to have the most information that is not present in the other variables.

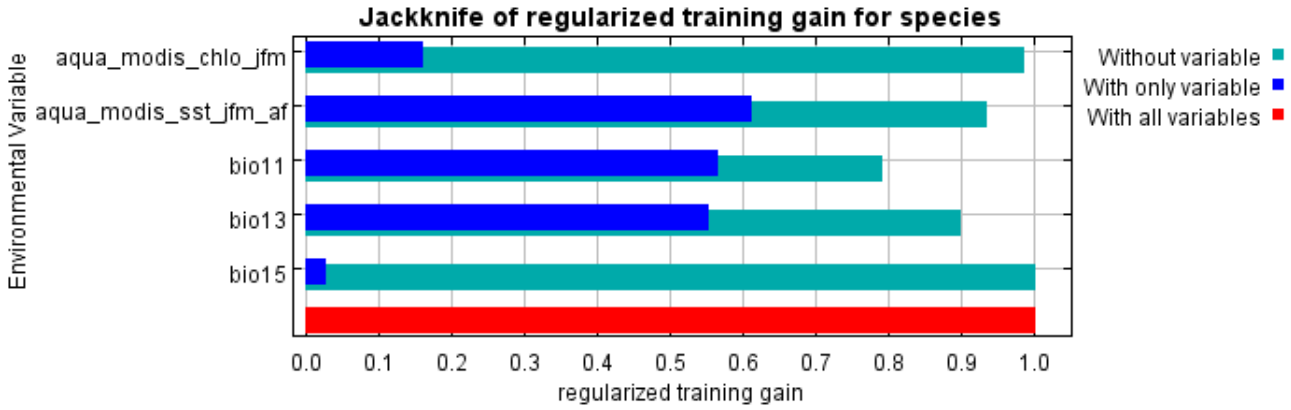

Figure 5.6: Jackknife test from MaxEnt on the winter dataset.

5.2 Response Curves

The African penguins' distribution is negatively influenced by an increase in the sea surface temperature, land surface temperature and percipitation. SST is the highest contributing factor, whilst annual mean land temperature has the highest permutation importance. There is a scaling factor of 10 for the temperature. A small increase causes a disproportionally large decline in penguin presence (for example with sensitivity analysis, from 7 degrees Celsius upwards for the annual trend). The outputs indicate that all probability of penguin presence are negatively correlated with their corresponding environmental variable, as can be seen in Figures 5.7 to 5.21. The y-axis shows the logistic output, probability of penguin presence.

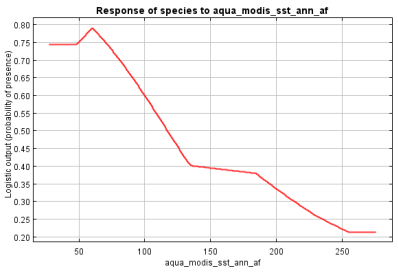

Figure 5.7: Response curve for annual sea surface temperature (degrees Celcius, x10).

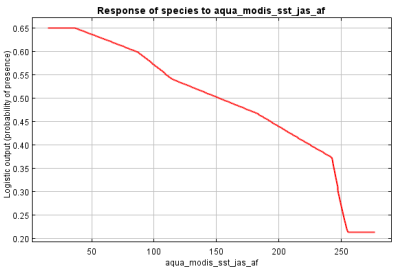

Figure 5.8: Response curve for summer sea surface temperature (degrees Celcius, x10).

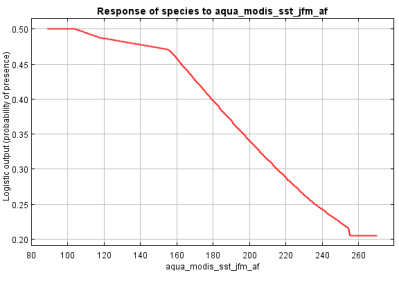

Figure 5.9: Response curve for winter sea surface temperature (degrees Celcius, x10).

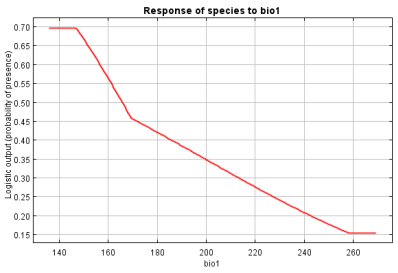

Figure 5.10: Response curve for annual mean temperature (degrees Celcius, x10).

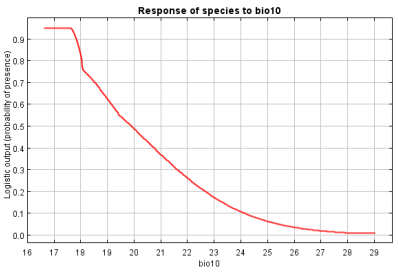

Figure 5.11: Response curve for mean temperature of warmest quarter.

Figure 5.12: Response curve for mean temperature of coldest quarter.

Figure 5.13: Response curve for annual percipitation (mm).

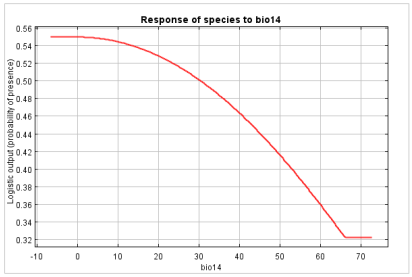

Figure 5.14: Response curve for summer percipitation.

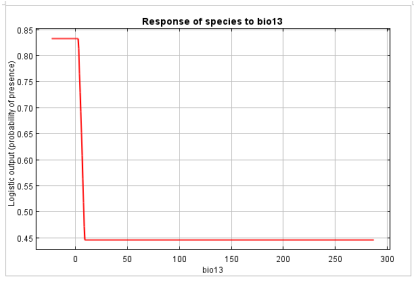

Figure 5.15: Response curve for winter percipitation.

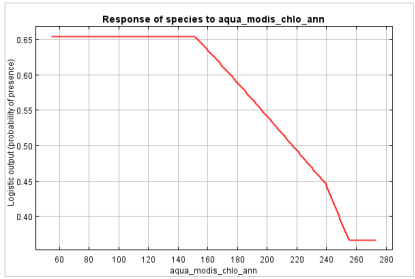

Figure 5.16: Response curve for annual chlorophyll count.

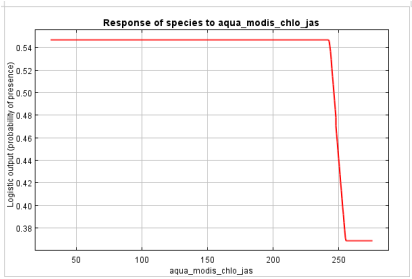

Figure 5.17: Response curve for summer chlorophyll count.

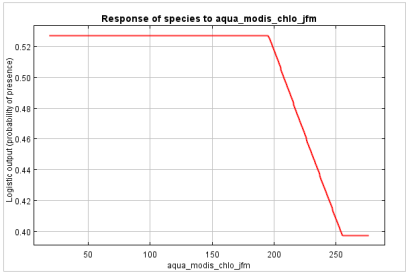

Figure 5.18: Response curve for winter chlorophyll count.

Figure 5.19: Response curve for annual percipitation coefficient of variation.

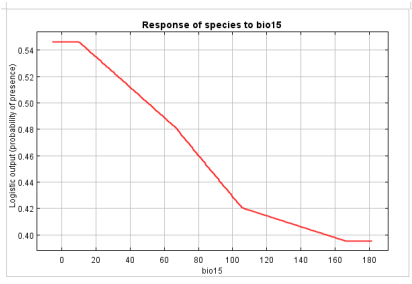

Figure 5.20: Response curve for summer percipitation coefficient of variation.

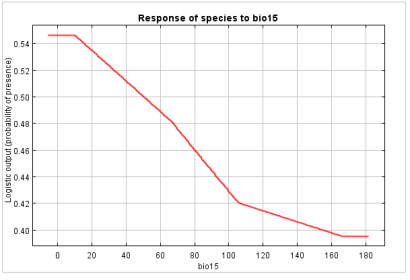

Figure 5.21: Response curve for winter percipitation coefficient of variation.

5.3 Suitability Mapping

MaxEnt version 3.3.3k, was applied to develop the SDMs for African penguins and it will be used to make projections for future restoration scenarios. MaxEnt shows good performance for presence-only data modelling methods. It fits complex responses to indicate the probability of the penguins' presence. The MaxEnt algorithm uses the environmental variables that are ecologically relevant to the species, to calculate a habitat suitability index, indicating where the species is most likely to occur. The output format is a logistic regression with continuous values from 0 to 1. In my SDM outputs, green is the least suitable and red is the most suitable areas (1).

```
#Project the model to geographical space:
envStack2 = envStack
projIDs = unique(c(pres$ID, pres$IDland, back$ID, back$IDland))
envStack2[!rasID@data@values_in_projIDs]  = NA

#Creates a raster that is just shows coastal grid cells
projMx = predict(envStack2, mxMod, na.rm=T)
projMx_P = rasterToPoints(projMx)
write.csv(projMx_P, file="projMx.csv")

plot(projMx, useRaster = TRUE, col=colorRampPalette(c("darkred", "red3",
"orange2", "orange", "yellow", "lightskyblue","steelblue3", "royalblue3",
"darkblue"))(12), cex= 2)

#c("yellow", "orange", "orange2", "red3", "darkred"))(12), cex = 1.5 )
 points(p$x,p$y,col= 'darkblue' , pch= '*')#col='#00000048'
writeRaster(projMx, filename="Projected.tif",overwrite=T,RAT=F)

#change raster to spatial line for better display
setwd("C:/Users/Frieda/Desktop/Penguins/")
library(raster)
prjMx <- raster("Projected.tif")
y <- as(rasterToPolygons(clump(projMx>0.002886203), dissolve=FALSE), 'SpatialLines')
y <- as(rasterToPolygons(clump(projMx>0.002886203), dissolve=TRUE), 'SpatialLines')
plot(y, col=2, lwd= 2)
```

The suitability map for annual output that I obtained in R statistical programming, is shown in Figure 5.22. However, I used the suitability data in the mathematical tool called Mathematica (Appendix B), to obtain better quality output suitability maps. GIS could also be used by opening the projected .tif file.

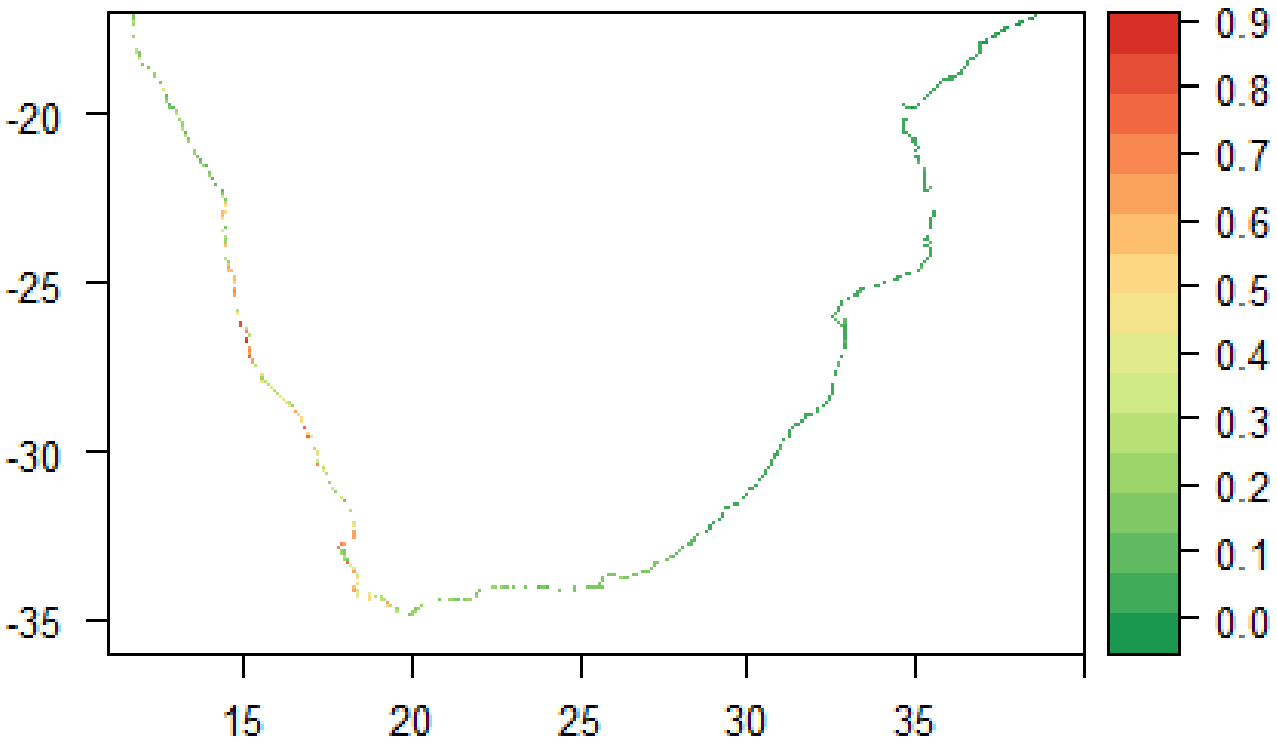

Figure 5.22: Annual suitability map obtained from R.

The SDM outputs are indicated in Figures 5.23 to 5.25. The ensemble predictions showed a continuous potential distribution for penguins in South Africa and Namibia. All the model outputs had a high discriminative ability, high AUC, as discussed in Chapter 4.6.

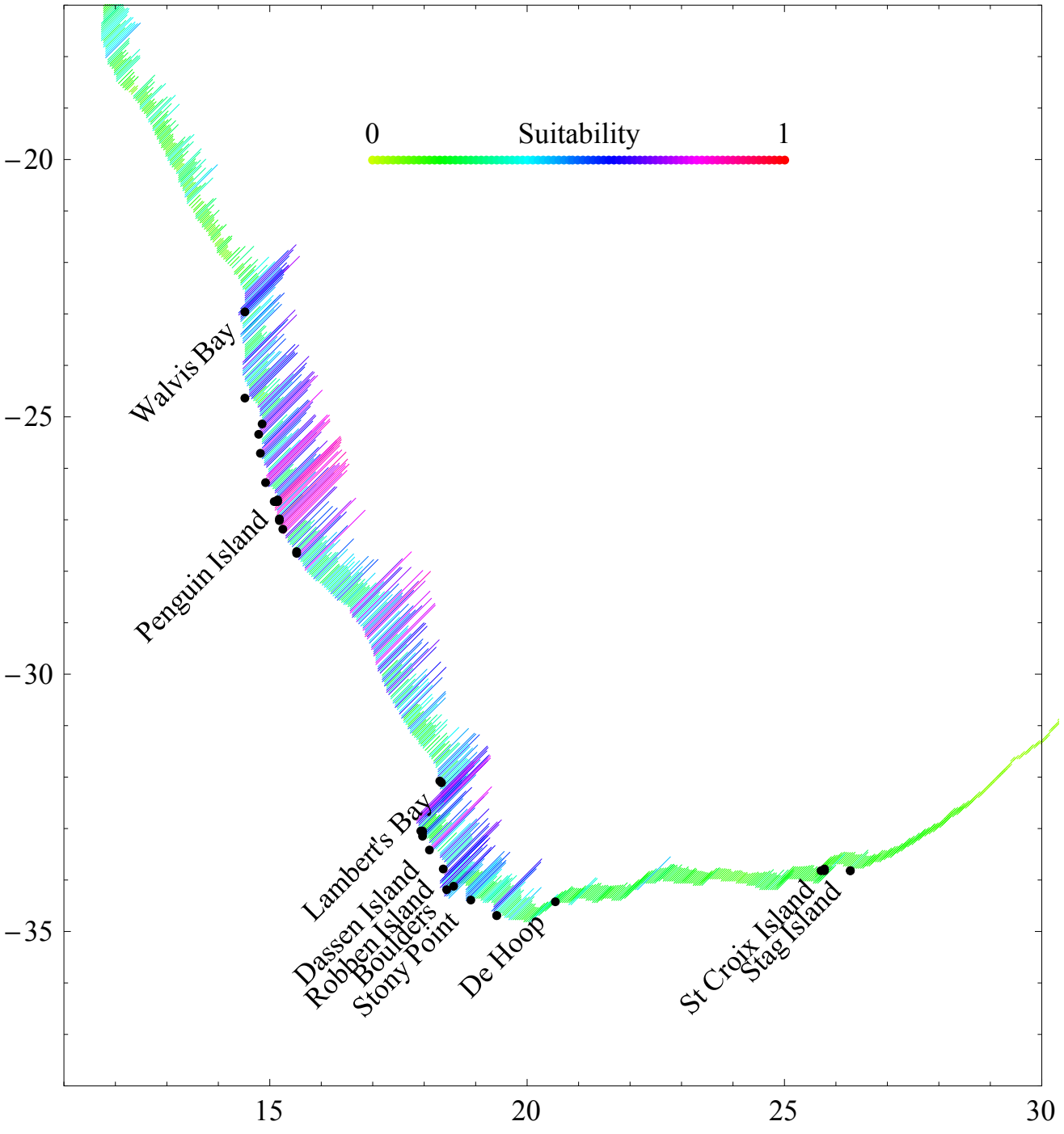

Figure 5.23: Annual SDM output indicating the lowest to highest suitability. The highest suitability indicates around regions near Penguin Island, while the lowest is St. Croix Island.

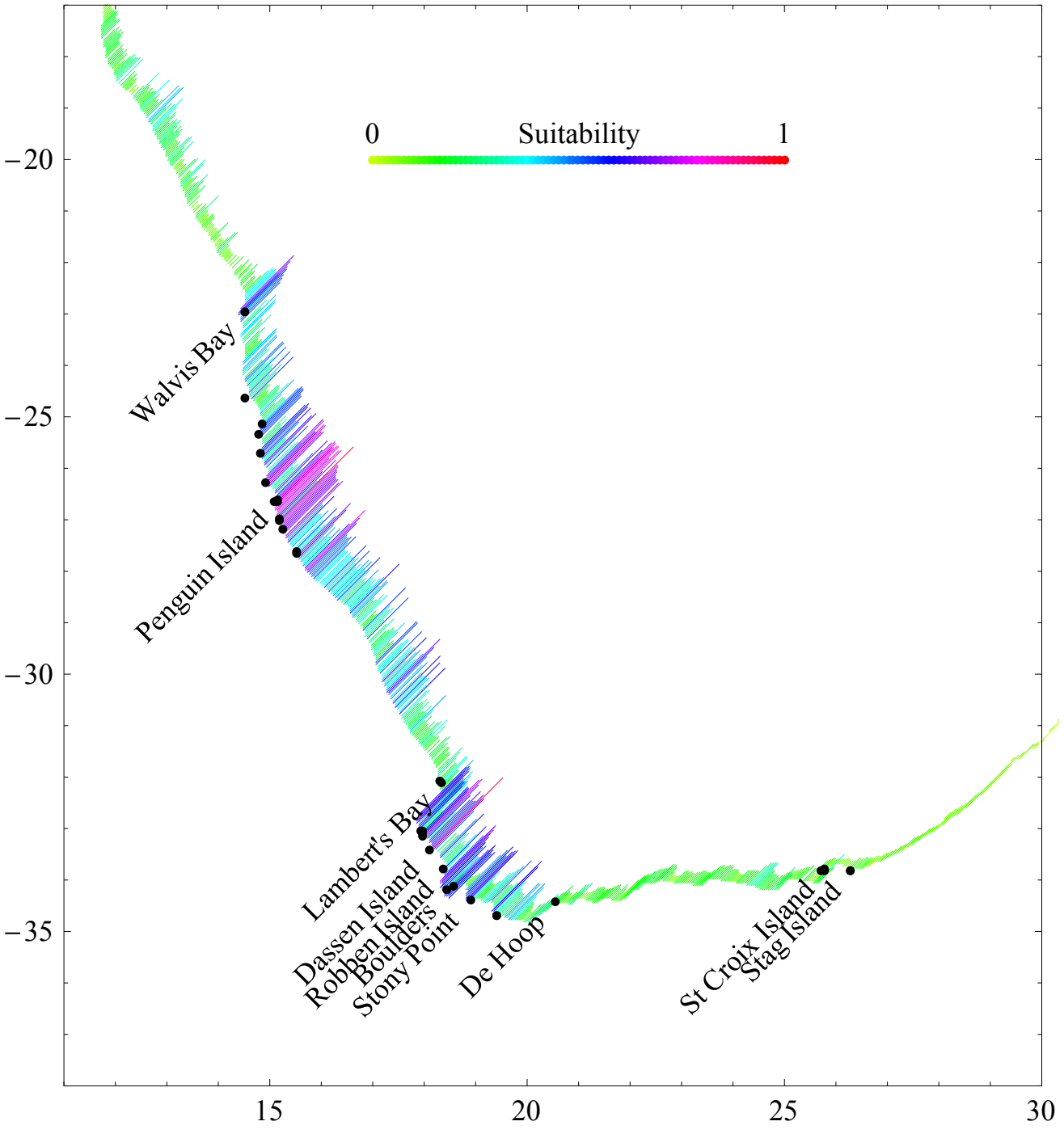

Figure 5.24: The SDM for summer.

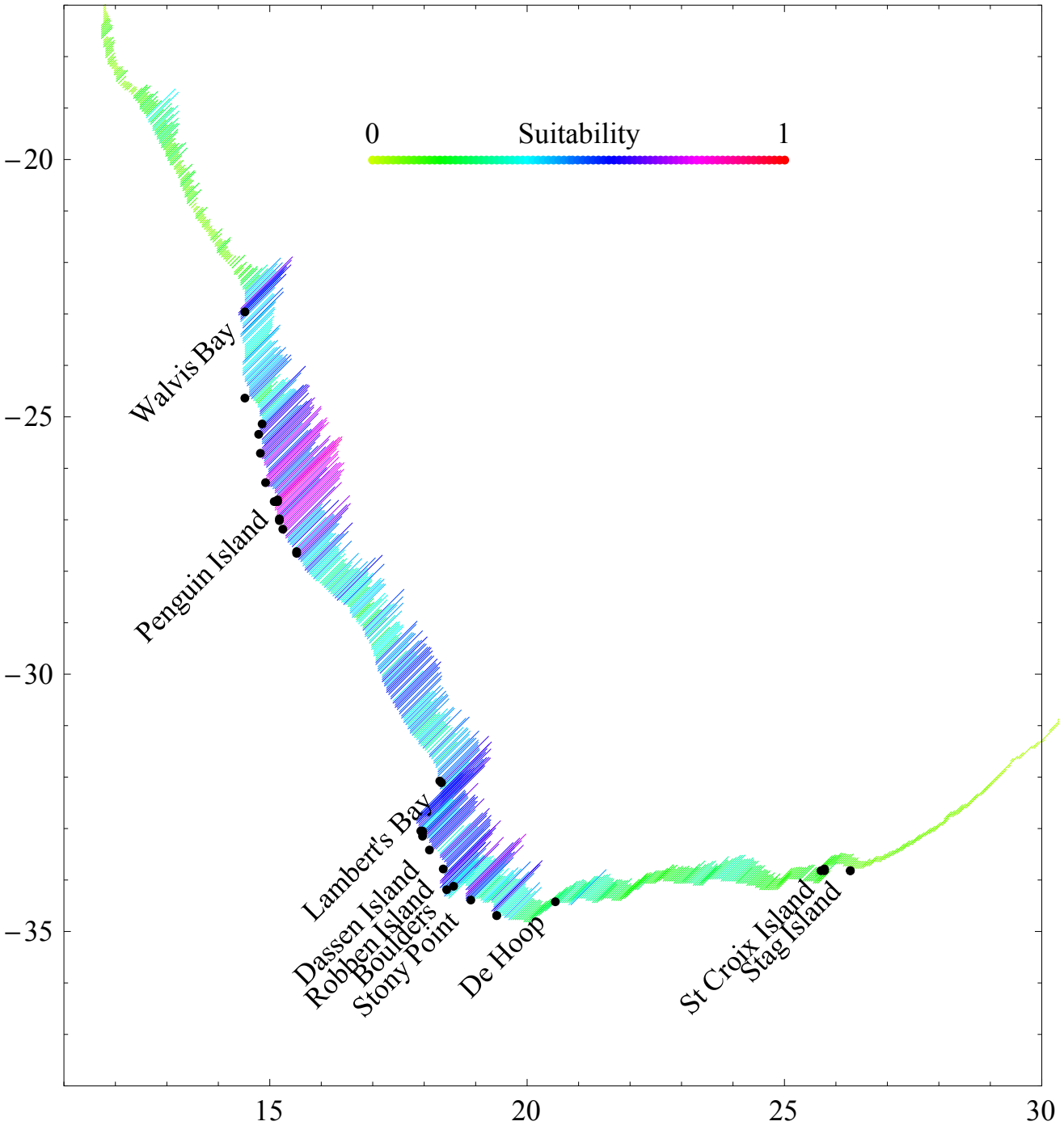

Figure 5.25: The SDM for winter.

Where the biggest colony is situated (St. Croix Island), the SDM indicates a low habitat suitability area, with more suitable habitat towards the Western Cape or Namibia. For the winter SDM output (Figure 5.25), it shows a bit more suitability in the area where St. Croix Island occurs, as SST is less of a contributing factor (46.9% rather than 72.4%).

5.3.1 Results Interpretation: Discussion of the Discrepancy Between the Model Predictions and the Actual Distribution of Penguins in the Region

There are islands along the coast (Western Cape and Algoa Bay, Eastern Cape) and two mainland sites that support the African penguin along the Western Cape coastline (see Figure 1.1). The mainland sites in South Africa include Stony Point, and Boulders Beach. The reason mainland colonies do well, is because the towns or human settlements function as barriers, keeping predators away from the penguins in much the same way as an island protects them. When one looks into why penguins occur where they do, and looks into where to place new colonies, it is important to see if there are suitable, rocky areas and predator-free sites such as offshore islands for them to establish. There are diamond activities taking place in the Northern Cape and Namibia, close to the coast, which disturb penguins, and this is a possible reason why they do not occur here in the first place.

I would like to clarify that my model prediction is based on environmental variables, as explained in the MaxEnt modelling chapter. The MaxEnt output format estimates the occurrence of the African penguin colonies, based on the environmental variables specified, thus it only estimates relative suitability. My model shows and is constrained to suitable environmental prediction according to the chosen environmental variables, but does not consider other factors as will be explained. The reason why few species in actual fact occupy all areas that satisfy their niche requirements are due to many possible factors, such as geographic barriers to dispersal, biotic interactions, and human modification of the environment. The MaxEnt model predicts colonies according to the environmental variables specified, thus the habitat suitable areas. The model predicts habitat suitable areas for colonies between Namibia and Western Cape along the west coast, even though they have not been measured there, possibly because of the diamond activities mentioned before.

To further explain, when the total population size is known, models predict the occurrence rate in a cell, defined as the expected number of colonies in that cell (Fithian and Hastie 2012). However, population size is usually unknown, such as in my current model, and only relative comparisons among these rates can be performed, resulting in a relative occurrence rate (ROR). Given that a colony was observed, the ROR describes the relative probability that the colony derived from each cell on the landscape. In other words, the ROR is the relative probability that a cell is contained in a collection of presence samples. The ROR corresponds to MaxEnt's raw output. MaxEnt can be used to predict the probability of presence by using a transformation of the ROR, called logistic output (Phillips and Dudík 2008), which relies on strong assumptions (Royle et al. 2012). We can interpret the MaxEnt model's predictions as indices of habitat suitability.

My results show system-wide ecological traps or scenarios in which organisms settle in habitats of poor quality, due to changes in stocks of particular foods due to over-fishing, or underlying environmental change. For example, although the Port Elisabeth (PE) area shows a low habitat suitability, where most of the African penguins occur, their prey, such as sardine and anchovy, is still found in the PE area. Abundant supplies of the penguin's favoured prey, such as sardine and anchovy, are no

longer present where the penguins expect to find them.

From Chapter 4.5, the most recent stock assessment from DAFF indicates, from their 2016 survey, that more than half the percentage of sardine and anchovy left are found to the west of Cape Agulhas. A big proportion of the African penguins occur in this area.

When establishing new African penguin colonies many factors need to be taken into account. The colony based SDM shows a poor predictor of where penguins are, as the SDM takes into consideration only presence data and environmental variables, where many other factors are applicable. I do believe that that MaxEnt modelling method gives a thorough, credible result on the situation it is based on. Using the MaxEnt technique, based on maximising entropy, under certain constraints, as explained, gives a reliable prediction according to what it takes into consideration.

Chapter 6

Discussion and Conclusion

6.1 Discussion

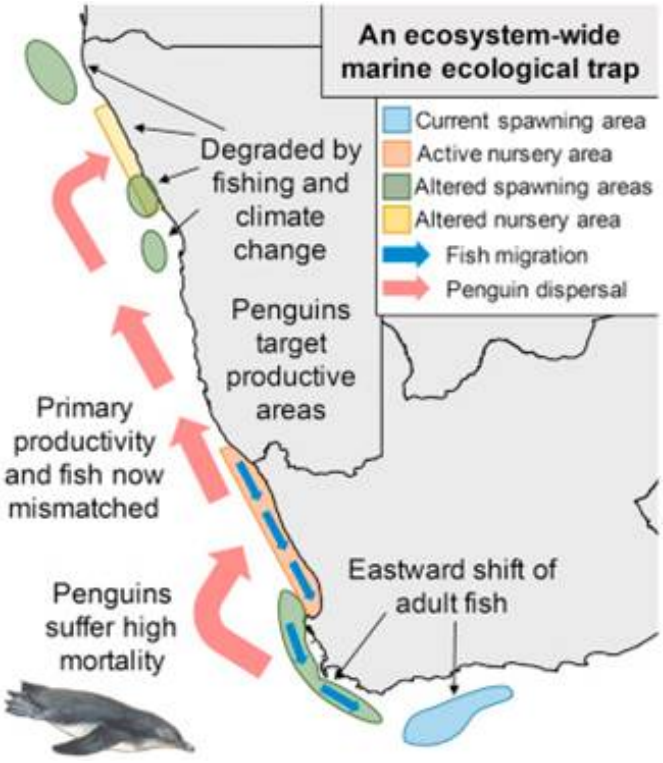

Figure 6.1: System Wide Ecological trap for African penguins (Source: Sherley et al.)

Scientists suspect that climate change may be causing sea surface temperatures (which is the main contributing variable to the SDM models in this study) to rise. This impacts the abundance of the African penguins' prey, by causing a shift in the prey distribution to locations beyond the historic breeding range of the penguins.

The adults must thus travel farther, and expend more energy, in order to find adequate food for

themselves and their chicks. The result is an increase in penguin mortality, due to starvation.

It is worthwhile to also mention the following. It has been shown that when young African penguins leave their nests for the first time, they do so alone. They have no guidance from their parents, and have to use their instinct to follow cues in their environment, both to find food and to stay alive in their first months at sea. Penguins utilize thermoclines that front cool waters, as a potential cue to the presence to prey. Abundant supplies of their favoured prey, such as sardine and anchovy, are no longer present where the penguins expect to find them. This causes the young birds to fall into what is known as an ecological trap. This is when they follow the usual cues to feeding grounds, only to find that the sources of food in these places are no longer available. The birds (measured using satellite transmitters) travelled large distances to areas of the ocean where sea temperatures were low, and chlorophyll concentrations were high, that have been historically known for their high fish abundance. The productivity, in the form of the microscopic phytoplankton food, which is the base of many aquatic food webs, was however high, although prey was low. This could be due to changes in stocks of particular foods due to over-fishing or underlying environmental change. Since the lower levels of the ecosystem have not been affected in the same way, the signals that the penguins would have previously used to locate their prey, are still intact. For example, the phytoplankton is still there and is still preyed upon by zooplankton. These are microscopic animals drifting in the ocean. Today, however, the fish that would normally co-occur with their plankton prey, are scarce or absent. Juvenile penguins are "tricked" into selecting the now poor habitat, and fall into this large-scale ecological trap. These previously unnoticed ecosystem-wide phenomena, explain the low survival chances of this endangered species, especially during its first year at sea. This contributes to the large decline of the African penguin population.

6.2 Conclusion

We projected the suitability maps of the African penguin's distributions using SDMs. The largest percentage of current colonies (about 44% of South Africa's penguins are in Algoa Bay) are located in suboptimal habitats. Given the rapid pace of environmental change, the scenario of an ecological trap, where organisms prefer to settle in poor-quality habitats, occurs. The SDMs address this issue. The importance of SDMs has been highlighted, and the output described acts as a baseline assessment.

The demography of the African penguin has been investigated. According to the MaxEnt output analysis, considering various environmental variables, sea surface temperature is the largest environmental contributor, with 72.4% for the annual factors, 53.2% for the summer factors and 46.9% for winter factors, towards the habitat suitability of the African penguin. Mean land temperature is the second biggest contributor for all seasons investigated.

The results inform the decision making process, by guiding environmental policies and land use planning, and by assisting in species' conservation.

Boulders area is an example of where from just 2 breeding pairs in 1982, the penguin colony has

grown to about 3 000 birds in recent years. This is partly due to the reduction in commercial pelagic trawling in False Bay. This has increased the supply of pilchards and anchovy, which form part of the penguins' diet.

In terms of the most suitable areas, using the environmental data explained, and as shown in Table 6.1 (or SDM outputs), Namibian areas around Penguin Island seem to be the most suitable. However, the prey availability has not been taken into account. Possible areas to relocate or establish colonies under investigation is the old De Hoop colony (which went extinct in 2006) and a site near Plettenberg Bay (which would be a completely new site), according to BirdLife. Camera traps for checking predators have been placed since November 2016. Establishing new African penguin colonies is an ongoing investigation with the shifting distribution of their prey.

Table 6.1: Most Suitable Locations for African Penguins According to MaxEnt Output.

x	y	Suitability Layer
15.1875	-26.9792	0.8609
15.1042	-26.7292	0.8565
15.2292	-27.1042	0.8535
15.1042	-26.6875	0.8514
15.1042	-26.7708	0.8506
15.2708	-27.1875	0.8444
16.8542	-29.2708	0.8438
15.1042	-26.6458	0.8429
15.1042	-26.8125	0.8354
15.1458	-26.8958	0.8333
14.9375	-26.3125	0.8193
15.2708	-27.2708	0.8175
14.9375	-26.2708	0.8162
14.9375	-26.1875	0.8122
15.2708	-27.1042	0.8014
15.2708	-27.1458	0.7963
16.5625	-28.8125	0.7915
14.9375	-26.1458	0.7874
18.1042	-32.7708	0.7866
15.1458	-26.8542	0.7836

6.3 Future Recommendations

Competition for food between seabirds and fisheries should be controlled over a larger scale. Spatial management of fisheries is an urgent requirement to increase food availability for penguins. I plan to incorporate the fish data as species distribution models, modelling their biomass/numbers. Each

suitable fish species will be modelled over different time periods. Modelling fishing abundance can be done using the co-kriging or applied hierarchical modelling technique. Thereafter, one could do a buffer analysis, in order to calculate the mean fish biomass available to the penguin colonies, from any given coordinate on the coastline.

Adding additional environmental variables to the SDMs, such as the human footprint, the vegetation layer and incorporating islands, can be added in future.

In addition to MaxEnt, methods such as BRT and BIOCLIM could also be used. There are many mathematical modelling approaches that could be taken into consideration for a more comprehensive solution, such as modelling the meta-community dynamics of regional colonies. The understanding of the patterns of movement of the penguins, taking into consideration habitat suitability, can be modelled using connectivity / adjacency matrices. Density dependence and the overcrowding effect can be investigated. Stochastic gravity models can be used to quantify dispersal effects and allow us to use standard statistical inference tools, such as maximum likelihood estimation and model selection based on information criteria.

Constrained paths and their mappings, on both the home and foraging territories of the penguins, can be investigated. To model the data, linear algebra, differential equations, and complex analysis can be used.

For individual based models, adaptive learning and decision making processes can be used to further improve models. Already we note that African penguins resort to jellyfish and goby fish in the Namibia area because of unavailability of their preferred prey. Both utility theory (rational) and prospective theory (irrational, based on past experience), can be investigated. To facilitate prediction, we can use applicable covariates. Random walk and step-selection analysis can be used to deal with this problem of availability at scale. A simple extension to this approach, termed integrated step selection analysis, could also be used.

The future research can make use of a resource selection function. This is a model of the likelihood that an available spatial unit will be used by an animal, given its resource value. This research will incorporate these stochastic processes and statistical mechanics to accomplish the goals of modelling the African penguin colonies, and addressing the stipulated challenges.

Bibliography

[1] Omri Allouche, Asaf Tsoar, and Ronen Kadmon. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of Applied Ecology*, 43(6):1223–1232, 2006.

[2] Miguel B Araujo, Richard G Pearson, Wilfried Thuiller, and Markus Erhard. Validation of species–climate impact models under climate change. *Global Change Biology*, 11(9):1504–1513, 2005.

[3] Mike Austin. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, 200(1):1–19, 2007.

[4] MP Austin. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157(2):101–118, 2002.

[5] Volker Bahn and Brian J McGill. Testing the predictive performance of distribution models. *Oikos*, 122(3):321–331, 2013.

[6] James Battin. When good animals love bad habitats: ecological traps and the conservation of animal populations. *Conservation Biology*, 18(6):1482–1491, 2004.

[7] Céline Bellard, Cleo Bertelsmeier, Paul Leadley, Wilfried Thuiller, and Franck Courchamp. Impacts of climate change on the future of biodiversity. *Ecology Letters*, 15(4):365–377, 2012.

[8] Lluís Brotons, Wilfried Thuiller, Miguel B Araújo, and Alexandre H Hirzel. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 27(4):437–448, 2004.

[9] Gemma Carroll, Jason D Everett, Robert Harcourt, David Slip, and Ian Jonsen. High sea surface temperatures driven by a strengthening current reduce foraging success by penguins. *Scientific Reports*, 6:22236, 2016.

[10] Maëlle Connan, GJ Greg Hofmeyr, and Pierre A Pistorius. Reappraisal of the trophic ecology of one of the world's most threatened spheniscids, the african penguin. *PloS One*, 11(7):e0159402, 2016.

[11] RJM Crawford, R Altwegg, BJ Barham, PJ Barham, JM Durant, BM Dyer, D Geldenhuys, AB Makhado, L Pichegru, PG Ryan, et al. Collapse of south africa's penguins in the early 21st century. *African Journal of Marine Science*, 33(1):139–156, 2011.

[12] Robert JM Crawford, Philip A Whittington, A Paul Martin, AJ Tree, and AB Makhado. Population trends of seabirds breeding in south africa's eastern cape, and the possible influence of anthropogenic and environmental change. *Marine Ornithology*, 37:159–174, 2009.

[13] RJM Crawford, JHM David, LJ Shannon, J Kemper, NTW Klages, JP Roux, LG Underhill, VL Ward, AJ Williams, and AC Wolfaardt. African penguins as predators and prey-coping (or not) with change. *African Journal of Marine Science*, 23:435–447, 2001.

[14] Robert JM Crawford, Peter J Barham, Les G Underhill, Lynne J Shannon, Janet C Coetzee, Bruce M Dyer, T Mario Leshoro, and Leshia Upfold. The influence of food availability on breeding success of african penguins spheniscus demersus at robben island, south africa. *Biological Conservation*, 132(1):119–125, 2006.

[15] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[16] Jane Elith, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. A statistical explanation of maxent for ecologists. *Diversity and Distributions*, 17(1):43–57, 2011.

[17] Jane Elith, Michael Kearney, and Steven Phillips. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4):330–342, 2010.

[18] Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.

[19] Jane Elith and Catherine H Graham. Do they? how do they? why do they differ? on finding reasons for differing performances of species distribution models. *Ecography*, 32(1):66–77, 2009.

[20] Stephen E Fick and Robert J Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 2017.

[21] Stephen E Fienberg, Alessandro Rinaldo, et al. Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40(2):996–1023, 2012.

[22] Janet Franklin. *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, 2010.

[23] Janet Franklin. Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, 16(3):321–330, 2010.

[24] Henrik Gislason, Michael Sinclair, Keith Sainsbury, and Robert O'boyle. Symposium overview: incorporating ecosystem objectives within fisheries management. *ICES Journal of Marine Science*, 57(3):468–475, 2000.

[25] Antoine Guisan, Reid Tingley, John B Baumgartner, Ilona Naujokaitis-Lewis, Patricia R Sutcliffe, Ayesha IT Tulloch, Tracey J Regan, Lluis Brotons, Eve McDonald-Madden, Chrystal Mantyka-Pringle, et al. Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12):1424–1435, 2013.

[26] Antoine Guisan and Wilfried Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009, 2005.

[27] Antoine Guisan and Niklaus E Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2):147–186, 2000.

[28] Robert J Hijmans and Jane Elith. Species distribution modeling with r, available online: http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf. *R Package Version 0.8-11*, 2013.

[29] Robert J Hijmans, Susan E Cameron, Juan L Parra, Peter G Jones, and Andy Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978, 2005.

[30] R Lewison, Daniel Oro, B Godley, L Underhill, Stuart Bearhop, R Wilson, D Ainley, José Manuel Arcos Pros, PD Boersma, P Borboroglu, et al. Research priorities for seabirds: improving seabird conservation and management in the 21st century. *Endangered Species Research, 2012, vol. 17, num. 2, p. 93-121*, 2012.

[31] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008.

[32] Richard H Moss, Jae A Edmonds, Kathy A Hibbard, Martin R Manning, Steven K Rose, Detlef P Van Vuuren, Timothy R Carter, Seita Emori, Mikiko Kainuma, Tom Kram, et al. The next generation of scenarios for climate change research and assessment. *Nature*, 463(7282):747, 2010.

[33] Pedro P Olea, Patricia Mateo-Tomás, and Ángel De Frutos. Estimating and modelling bias of the hierarchical partitioning public-domain software: implications in environmental management and conservation. *PLoS One*, 5(7):e11698, 2010.

[34] Michela Pacifici, Wendy B Foden, Piero Visconti, James EM Watson, Stuart HM Butchart, Kit M Kovacs, Brett R Scheffers, David G Hole, Tara G Martin, H Resit Akcakaya, et al. Assessing species vulnerability to climate change. *Nature Climate Change*, 5(3):215, 2015.

[35] Richard G Pearson, Terence P Dawson, and Canran Liu. Modelling species distributions in britain: a hierarchical integration of climate and land-cover data. *Ecography*, 27(3):285–298, 2004.

[36] A Townsend Peterson, Monica Papeş, and Jorge Soberón. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213(1):63–72, 2008.

[37] Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197, 2009.

[38] S Phillips. A brief tutorial on maxent. at&t res, 2008.

[39] Steven J Phillips and Miroslav Dudík. Modelling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175, 2008.

[40] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259, 2006.

[41] RM Randall. Jackass penguins. *Oceans of life off southern Africa*, 244:256, 1989.

[42] Óscar Rodríguez de Rivera and Antonio López-Quílez. Development and comparison of species distribution models for forest inventories. *ISPRS International Journal of Geo-Information*, 6(6):176, 2017.

[43] William Michael Lewin Robinson. Modelling the impact of the south african small pelagic fishery on african penguin dynamics. *University of Cape Town*, 2013.

[44] Richard B Sherley, Katrin Ludynia, Bruce M Dyer, Tarron Lamont, Azwianewi B Makhado, Jean-Paul Roux, Kylie L Scales, Les G Underhill, and Stephen C Votier. Metapopulation tracking juvenile penguins reveals an ecosystem-wide ecological trap. *Current Biology*, 27(4):563–568, 2017.

[45] Richard B Sherley, Henning Winker, Res Altwegg, Carl D van der Lingen, Stephen C Votier, and Robert JM Crawford. Bottom-up effects of a no-take zone on endangered penguin demographics. *Biology Letters*, 11(7):20150237, 2015.

[46] Phil N Trathan, Pablo García-Borboroglu, Dee Boersma, Charles-André Bost, Robert JM Crawford, Glenn T Crossin, Richard J Cuthbert, Peter Dann, Lloyd Spencer Davis, Santiago De La Puente, et al. Pollution, habitat loss, fishing, and climate change as critical threats to penguins. *Conservation Biology*, 29(1):31–41, 2015.

[47] LG Underhill, RJM Crawford, AC Wolfaardt, PA Whittington, BM Dyer, TM Leshoro, M Ruthenberg, L Upfold, and J Visagie. Regionally coherent trends in colonies of african penguins spheniscus demersus in the western cape, south africa, 1987–2005. *African Journal of Marine Science*, 28(3-4):697–704, 2006.

[48] Jeremy VanDerWal, Luke P Shoo, Catherine Graham, and Stephen E Williams. Selecting pseudo-absence data for presence-only distribution modelling: how far should you stray from what you know? *Ecological Modelling*, 220(4):589–594, 2009.

[49] Carl Walters and Peter H Pearse. Stock information requirements for quota management systems in commercial fisheries. *Reviews in Fish Biology and Fisheries*, 6(1):21–42, 1996.

[50] Florian Weller, Lee-Anne Cecchini, Lynne Shannon, Richard B Sherley, Robert JM Crawford, Res Altwegg, Leanne Scott, Theodor Stewart, and Astrid Jarre. A system dynamics approach to modelling multiple drivers of the african penguin population on robben island, south africa. *Ecological Modelling*, 277:38–56, 2014.

[51] Robert P Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J Hijmans, Falk Huettmann, John R Leathwick, Anthony Lehmann, Jin Li, Lucia G Lohmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

Appendix A: R Code

```
#loading packages
library(rasterVis)
library(biogeo)
library(dismo)
library(raster)
library(sp)
library(spatial.tools)
library(corrplot)
library(rgdal)
#Set working directory:
setwd("C:/Users/Frieda/Desktop/Penguins/")

#Read in presence data:
pres = read.csv('Penguin localities_Frieda.csv')

#Get shapefile with coastline:
coast = shapefile('GIS/GSHHS_f_L1_Africa.shp')
coastCoords = coast@polygons[[1]]@Polygons[[1]]@coords
coastCoords = data.frame(coastCoords)
names(coastCoords) = c('x','y')
coastCoords = coastCoords[coastCoords$x>11 & coastCoords$x<40
& coastCoords$y>c(-36) & coastCoords$y<c(-17),]

#Get environmental layers (rasters):
envFiles <- list.files("GIS/") #Names of all files in GIS directory
envFiles <- envFiles[c(grep('jpg', envFiles), grep('tif', envFiles))]
#Select only jpg and tif files
envFiles = envFiles[-grep('xml',envFiles)] #Deselect xml files
envNames = {} #Vector to store names of rasters
#Loop through raster file names, open, crop, assign and place in raster stack:
for(r in 1:length(envFiles)){
  rast <- raster(paste("GIS/",envFiles[r],sep=""), RAT=F)
  rast = crop(rast, extent(c(11,40,-36,-17)))
  assign(names(rast), rast)
  envNames[r] = names(rast)
  if(r==1){
    envStack = stack(rast)
  } else if(r>1 & r<=10){
```

```
    envStack = stack(envStack, rast)
  }
  rm(rast) #Remove placeholder raster
}
#Plotting raster rayers
#plot(envStack)
#Rasters do not completely overlap.
#Use fix_extent to make them align and add to raster stack:

envStack2 = fix_extent(bio1, envStack)
envStack = addLayer(envStack2[[1]], bio1)
envStack2 = fix_extent(bio5, envStack)
envStack = addLayer(envStack2[[1]], bio5)
envStack2 = fix_extent(bio12, envStack)
envStack = addLayer(envStack2[[1]], bio12)
envStack2 = fix_extent(bio15, envStack)
envStack = addLayer(envStack2[[1]], bio15)
#envStack2 = fix_extent(bio11, envStack)
#envStack = addLayer(envStack2[[1]], bio11)
#envStack2 = fix_extent(bio13, envStack)
#envStack = addLayer(envStack2[[1]], bio13)
#envStack2 = fix_extent(bio14, envStack)
#envStack = addLayer(envStack2[[1]], bio14)

#Move points on land to nearest cell in sea AND points in sea to land:
#Add fields required by biogeo:
checkdatastr(pres)
pres$Species = 'Penguin'
pres = addmainfields(pres, species = 'Species')
head(pres)
write.csv(pres, file = "pres_v1.csv")
#Move points on land to nearest cell in sea

#presLand = nearestcell(pres, bio1)
presLand = pres

#renaming the x and y variables
# colnames(presLand)[2] ="x_land"
# colnames(presLand)[3] ="y_land"
# head(presLand)
```

```
#Create new columns for "land" coordinates:
pres$x_land = pres$x
pres$y_land = pres$y

#pres$x_land[pres$ID%in%presLand$moved$ID] = presLand$moved$x
#pres$y_land[pres$ID%in%presLand$moved$ID] = presLand$moved$y
#There are two points that were not successfully moved to nearest land point. Move these:
# pres[is.na(pres$bio1),]
# pres$x_land[pres$location=='Seal Island Algoa Bay'] = 26.2673
# pres$y_land[pres$location=='Seal Island Algoa Bay'] = -33.7770
# pres$x_land[pres$location=='Hollams Bird Island'] = 14.5978
# pres$y_land[pres$location=='Hollams Bird Island'] = -24.6146
#Move points in sea to nearest cell on land:
#Newaqua_modis_chlo_amj = fix_extent(bio1, aqua_modis_chlo_amj)
# presSea = nearestcell(pres,aqua_modis_chlo_amj)
#error, missing coord. ref. to aqua_modis_chlo_amj
#Do the same for all coastal coordinates (will use these as background points):
back = coastCoords
checkdatastr(back)
back$Species = 'Penguins'
back = addmainfields(back, species = 'Species')
head(back)

#Sea to land:
#back = read.csv("back.csv", header = TRUE, sep= ",")
# backLand = nearestcell(back, bio1) #na.action(na.omit(bio1(1,NA)))

# nrow(backLand$moved) #How many points moved

back$x_land = back$x
back$y_land = back$y
# rename the variables in back, x and y to x_land and y_land
head(back)
backLand = back

# back$x_land[back$ID%in%backLand$moved$ID] = backLand$moved$x
# back$y_land[back$ID%in%backLand$moved$ID] = backLand$moved$y
#remove two columns

#Land to sea:
# backSea = nearestcell(back, aqua_modis_chlo_amj)
```

```
#error, missing the coord. ref. to aqua_modis_chlo_amj

#Extract environmental data:
presEnv = data.frame(extract(envStack, cbind(pres$x, pres$y)))
presEnvLand = data.frame(extract(envStack, cbind(pres$x_land, pres$y_land)))
#write.csv(presEnvLand, file = "presEnvLand.csv")
#Replace any NA values (points in the sea) with values using land coordinates:
presEnv[is.na(presEnv)] = presEnvLand[which(is.na(presEnv), arr.ind=T)]
backEnv = data.frame(extract(envStack, cbind(back$x_land, back$y_land)))
# backEnv = data.frame(extract(envStack, cbind(back$x, back$y)))
backEnvLand = data.frame(extract(envStack, cbind(back$x_land, back$y_land)))
#write.csv(backEnvLand, file = "backEnvLand.csv")
#Replace any NA values (points in the sea) with values using land coordinates:
backEnv[is.na(backEnv)] = backEnvLand[which(is.na(backEnv), arr.ind=T)]

#Add environmental data to coordinate data:
pres = cbind(pres[,c('location','x','y','x_land','y_land')], presEnv)
back = cbind(back[,c('x_land','y_land','x_land','y_land')], backEnv)
#write.csv(pres, file = "pres_presEnv.csv")
#write.csv(back, file = "back_backEnvLand.csv")

#Remove duplicated points:
rasID = envStack[[1]]
rasID[] = 1:length(rasID)
writeRaster(rasID, 'rasID.tif', overwrite=T)
pres$ID = extract(rasID, pres[,c('x','y')])
pres$IDland = extract(rasID, pres[,c('x_land','y_land')])
back$ID = extract(rasID, back[,c('x_land','y_land')])
#back$ID = extract(rasID, back[,c('x','y')])
back$IDland = extract(rasID, back[,c('x_land','y_land')])

####
#pres = read.csv("pres_presEnv.csv")
pres = pres[!duplicated(pres$ID) & !duplicated(pres$IDland),]
back = back[!duplicated(back$ID) & !duplicated(back$IDland),]

#saving the pres and back # rm duplicated
#write.csv(pres, file = "pres_rm_dupl.csv")
#write.csv(back, file = "back_rm_dupl.csv")

#Remove "presence coordinates" from "back coordinates":
```

```
back = back[!back$ID%in%c(pres$ID,pres$IDland) & !back$IDland%in%c(pres$ID,pres$IDland),]

#write.csv(back, file= "back_rm_pres_coord.csv")

#Check for collinearity among environmental predictors:
envNames = names(envStack)
names(back)
colnames(back)[3] = "x"
colnames(back)[4]= "y"
names(back)

dat = rbind(pres[,-1],back[,names(back)%in%names(pres)])
M <- cor(dat[,envNames])

corrplot.mixed(M, lower = "circle", upper = "number", tl.pos = c("lt"))

p = pres[,c('x','y')]
a = back[,c('x','y')]
set.seed(10)
pSamp = sample(x = c(1:nrow(p)), size=round(nrow(p)*0.7,0), replace=F)
ptrain = p[pSamp,]
ptest = p[-pSamp,]
aSamp = sample(x = c(1:nrow(a)), size=round(nrow(a)*0.7,0), replace=F)
atrain = a[aSamp,]
atest = a[-aSamp,]

# The function uses environmental data for locations of known presence
and for a large number of 'background' locations.
Environmental data can be extracted from raster files.
The result is a model object that can be used to predict the suitability
of other locations, for example, to predict the entire range of a species.

#Full model with all occurrences:
betaRCs <- seq(0.05,0.95,0.05)
bestAUC = 10
names(envStack)
keeps = c('aqua_modis_chlo_ann','aqua_modis_sst_ann_af','bio1', 'bio12', 'bio15')
#Select which variables to keep
#mxMod = maxent(x = dat[,keeps],
#               p = c(rep(1,nrow(pres)), rep(0,nrow(back))),
#               args = c('-p','nothreshold',paste0('beta_lqp=',betaRCs[bestAUC]),
```

```
#                         paste0('beta_hinge=',betaRCs[bestAUC])),
#               removeDuplicates=T)
mxMod = maxent(x = dat[,keeps],
               p = c(rep(1,nrow(pres)), rep(0,nrow(back))),
               args = c('responsecurves','-p','nothreshold',
               paste0('beta_lqp=',betaRCs[bestAUC]),
                         paste0('beta_hinge=',betaRCs[bestAUC])),
               removeDuplicates=T)

#Analysis of variable contributions
#var_contr = c(65.3,15.3,5.8,4.6,4.1,3.1,1.5,0.2)
#var_mx = c('aqua_modis_sst_ann_af', 'bio1','aqua_modis_chlo_jas','aqua_modis_chlo_ann',
#'bio12','bio15','aqua_modis_chlo_ond','aqua_modis_chlo_jfm' )
#an_var = cbind(var_contr, var_mx)
#contr_var = data.frame(an_var)

#write.csv(contr_var, file = "an_var.csv")

#set up the plot size

#plot the hist in a pot
# plot(x=var_contr,  type="h", main="Analysis of variable contributions ",
 ylab= "Variable contribution ((%)",
#      col="dark red", lwd = 50, cex = 1, cex.lab =1.5, cex.main = 1)

#Project the model to geographical space:
envStack2 = envStack
projIDs = unique(c(pres$ID, pres$IDland, back$ID, back$IDland))
envStack2[!rasID@data@values%in%projIDs]  = NA
#Creates a raster that is just shows coastal grid cells
projMx = predict(envStack2, mxMod, na.rm=T)
projMx_P = rasterToPoints(projMx)
write.csv(projMx_P, file="mydata12July2017.csv")

plot(projMx, useRaster = TRUE, col=colorRampPalette(c("darkred", "red3",
"orange2", "orange", "yellow", "lightskyblue","steelblue3",

#c("yellow", "orange", "orange2", "red3", "darkred"))(12), cex = 1.5 )
 points(p$x,p$y,col= 'darkblue' , pch= '*')#col='#00000048'
writeRaster(projMx, filename="Projected.tif",overwrite=T,RAT=F)
```

```
#change raster to spatial line for better display
setwd("C:/Users/Frieda/Desktop/Penguins/")
library(raster)
prjMx <- raster("Projected.tif")
y <- as(rasterToPolygons(clump(projMx>0.002886203), dissolve=FALSE), 'SpatialLines')
y <- as(rasterToPolygons(clump(projMx>0.002886203), dissolve=TRUE), 'SpatialLines')
plot(y, col=2, lwd= 2)

###################
library(rasterVis)
levelplot(prjMx)
levelplot(prjMx, names.attr='One')
          #at=seq(0, 1, length.out=100),
          #par.strip.text=list(font=2, cex=1.2))


#mapTheme <- rasterTheme(region=rev(brewer.pal(8,"RdBu")))
mapTheme <- rasterTheme(region=(brewer.pal(8,"YlOrRd")))
plt <- levelplot(prjMx, margin=F, par.settings=mapTheme)
plt

mapTheme <- rasterTheme(region=rev(brewer.pal(8,"RdYlGn")))
plt <- levelplot(prjMx, margin=F, par.settings=mapTheme)
plt

################
#Run this to create subset plot without legend

jpeg(filename = "plt1.jpeg", width = 480, height = 480, units = "px",
     pointsize = 12,  quality = 75, bg = "white")
jpeg(filename = "plt1_default.jpeg")
plt1 <- levelplot(prjMx, margin=F, colorkey=F,
par.settings=mapTheme, xlim= c(10, 18), ylim=c(-35, -17))
plt1
dev.off()

jpeg(filename = "plt2.jpeg", width = 480, height = 480, units = "px",
     pointsize = 12,  quality = 75, bg = "white")
plt2 <- levelplot(prjMx, margin=F, colorkey=F, par.settings=mapTheme,
xlim= c(18, 30), ylim=c(-35, -32))
plt2
```

```
dev.off()

jpeg(filename = "plt3.jpeg", width = 480, height = 480, units = "px",
     pointsize = 12,  quality = 75, bg = "white")
plt3 <- levelplot(prjMx, margin=F, colorkey=F, par.settings=mapTheme,
xlim= c(30, 40), ylim=c(-32, -17))
plt3
dev.off()

############################
# x<- list(x= c(10, 18), y=c(-35, -17))
# plot(extent(x), add=T)
# p <- as(extent(x), 'SpatialPolygons')
# ?png
# plt + layer(p, col="gray", lwd=0.5)
# plt + layer(sp.polygons(p, pch=20, cex=0.1, col=2))

jpeg(filename = "wholeSA.jpeg", width = 1200, height = 600, units = "px",
     pointsize = 12,  quality = 75, bg = "white")
plt + layer(sp.polygons(p, pch=20, cex=0.1, col=2)) +
layer(sp.polygons(as(extent(list(x= c(18, 30), y=c(-35, -32))),
'SpatialPolygons'), pch=20, cex=0.1, col=2)) +
  layer(sp.polygons(as(extent(list(x= c(30, 40), y=c(-32, -17))),
  'SpatialPolygons'), pch=20, cex=0.1, col=2))
dev.off()
#text(SpatialPoints(10.16451, -34.11275), "A")

tiff(filename = "wholeSA.tiff", width = 800, height = 500, units = "px",
     pointsize = 1,  quality = 75, bg = "white")
plt + layer(sp.polygons(p, pch=20, cex=0.1, col=2)) +
  layer(sp.polygons(as(extent(list(x= c(18, 30), y=c(-35, -32))),
  'SpatialPolygons'), pch=20, cex=0.1, col=2)) +
  layer(sp.polygons(as(extent(list(x= c(30, 40), y=c(-32, -17))),
  'SpatialPolygons'), pch=20, cex=0.1, col=2))
dev.off()

############################################################
#See response curves:
response(mxMod)

#See Maxent results
```

```
mxMod@results

#install.packages("calibrate")
library(calibrate)
library(raster)
library(tiff)
tiff('Projected.tif')
ab = 'Projected.tif'
bc = 'raster(ab)'
cd=tiff(bc)

plot(cd,col=terrain.colors(10), lwd=5,xlim = c(10,40), ylim = c(-35,-17))
points(pres$x,pres$y )
textxy(pres$x, pres$y,pres$location)
dev.off()

library(calibrate)
library(raster)
pred = raster('Projected.tif')
plot(pred,col=terrain.colors(10), lwd=5)
points(pres$x,pres$y )
textxy(pres$x, pres$y,pres$location)

#####################################

pts <- SpatialPointsDataFrame(pres[, c("x", "y")], pres)

plt1 <- levelplot(prjMx, margin=F, par.settings=mapTheme, xlim= c(10, 18),
ylim=c(-35, -17)) +
  layer(sp.points(pts, pch=5, cex=0.5, col="black")) +
  layer(panel.text("location"))
  layer(panel.text(179500, 333000,
                text2add[panel.number()]))
plt1

  layer(sp.polygons(as(extent(list(x= c(18, 30), y=c(-35, -32))), 'SpatialPolygons')


  ####3#defining ....
projMx_P = round(projMx_P, digits = 4)
write.csv(projMx_P ,file ="projMx_P.csv")
```

```
A = data.frame(matrix(ncol= 3, nrow= 28))

colnames(A) = colnames(projMx_P)
  for(j in 1:28 ){
    for(i in 1: 1784){
       if (projMx_P$x[i]== pres$x[j] & projMx_P$y[i]== pres$y[j])
          {A[i,]=projMx_P[i,]}
  }

  }
```

Appendix B: Mathematica Code

```
a= Import [ "C:\\Users\\Frieda\\Desktop\\projMx_P.csv" , "CSV" ];

Dimensions [a]
\{1785,4\}

ListPlot [ Table [\{a[[i,2]],a[[i,3]]\},\{i,2,1785\}]]

Graphics [ Table [\{ Thin , Line [\{\{a[[i,2]],a[[i,3]]\},\{a[[i,2]]
+1.5*a[[i,4]],a[[i,3]]+1.5*a[[i,4]]\}\}]\},\{i,2,1785\}]];

Graphics [ Table [ Line [\{\{a[[i,2]],a[[i,3]]\},\{a[[i,2]]
+2*a[[i,4]],a[[i,3]]+2*a[[i,4]]\}\}],\{i,2,1785\}]];

tu1 = Graphics left [Table left [left lbrace Hue left [{j} / {100} right ],
Point left [left lbrace 15+0.1j,-20 right rbrace right ]
right rbrace , left lbrace j,20,100 right rbrace right ] right ] ;

tu2 = Graphics left [Table left [left lbrace Thin ,
Hue left [0.8a left [left [i,4 right ] right ] +0.2 right ] ,
Line left [left lbrace left lbrace a left [left [i,2 right ] right ] ,
a left [left [i,3 right ] right ] right rbrace , left lbrace a left
[left [i,2 right ] right ] +1.5*a left [left [i,4 right ] right ] ,
a left [left [i,3 right ] right ] +1.5*a left [left [i,4 right ] right ]
right rbrace right rbrace right ] right rbrace ,
left lbrace i,2,1785 right rbrace right ] right ] ;

b= Import [ "C:\\Users\\Frieda\\Desktop\\Penguin localities_Frieda2.csv" ];

tu3 = Graphics [ Table [\{ Point [\{b[[i,2]],b[[i,3]]\}]\},\{i,2,35\}]];

"(*" colorbar "*)"

tu4 = Graphics [ Table [\{ Text [b[[i,1]],\{b[[i,2]]-0.3,b[[i,3]]-0.3\},
 Right ,\{1,1\}]\},\{i,2,35\}]];

"(*" bar "*)"

tu5 = Graphics [\{ Text [ "0" ,\{17,-19.5\}],
Text [ "1" ,\{25,-19.5\}], Text [ "Suitability" ,\{21,-19.5\}]\}];
```

```
Show [ tu1 , tu2 , tu3 , tu4 , tu5 , PlotRange ->\{\{11,30\},\{-38,-17\}\}, Frame -> True ]
```